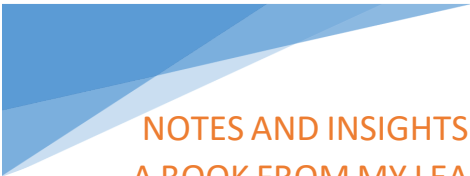




10/23/2020

Data Visualization



NOTES AND INSIGHTS TIED UP INTO
A BOOK FROM MY LEARNING IN THE
DATA VISUALIZATION FIELD



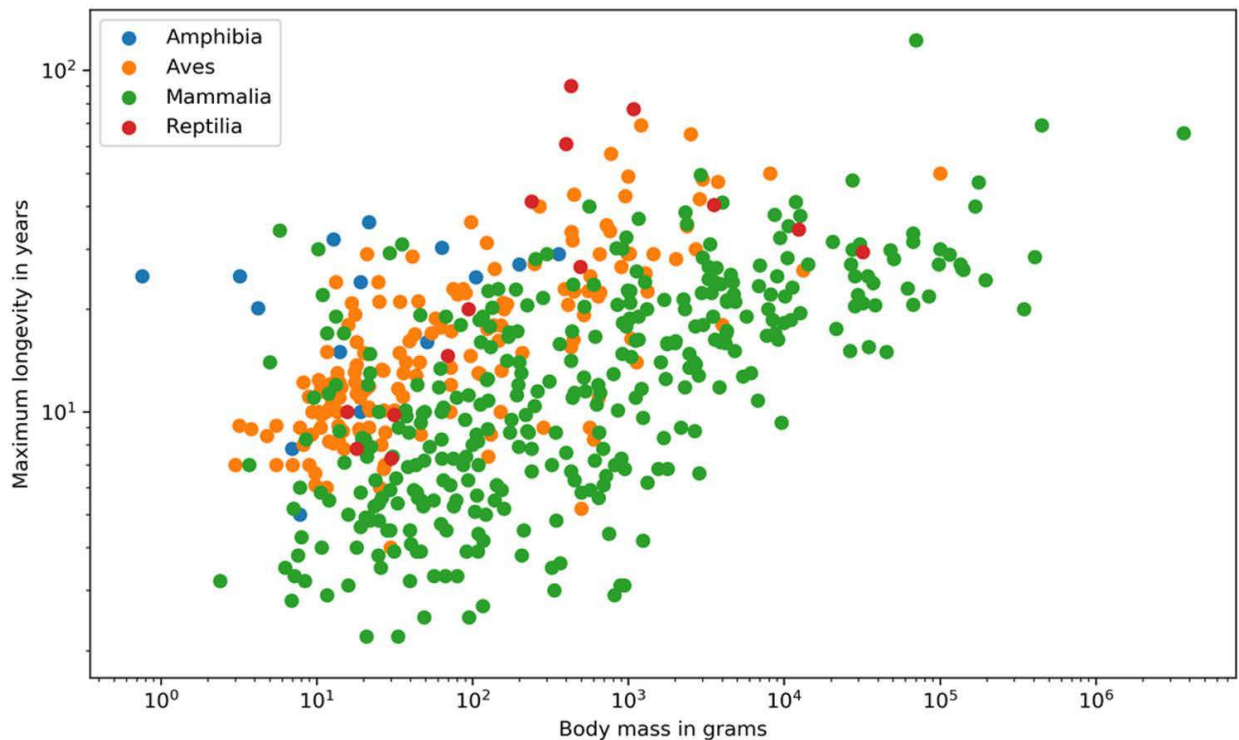
Sourabh S

Chapter 1: Importance of Data Visualization

Introduction:

Unlike machines, people are usually not equipped for interpreting a large amount of information from a random set of numbers and messages in each piece of data. Out of all our logical capabilities, we understand things best through the visual processing of information. When data is represented visually, the probability of understanding complex builds and numbers increases.

Instead of just looking at data in the columns of an Excel spreadsheet, we get a better idea of what our data contains by using visualization. For instance, 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 body mass and the maximum longevity of various animals grouped by class. There is a positive correlation between body mass and maximum longevity:



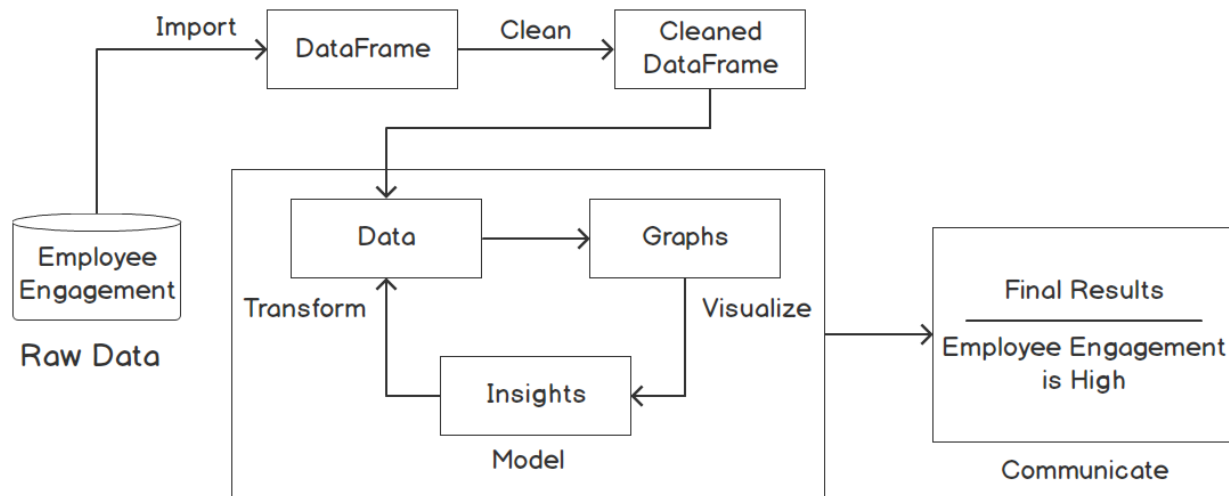
Visualizing data has many advantages, such as the following:

- Complex data can be easily understood.
- A simple visual representation of outliers, target audiences, and futures markets can be created.
- Storytelling can be done using dashboards and animations.
- Data can be explored through interactive visualizations.

Chapter 1: Importance of Data Visualization

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.



Overview of Statistics:

Measures of Central Tendency

Measures of central tendency are often called **averages** and describe central or typical values for a probability distribution.

- **Mean:** The arithmetic average is computed by summing up all measurements and dividing the sum by the number of observations.
- **Median:** This is the middle value of the ordered dataset. If there is an even number of observations, the median will be the average of the two middle values. The median is less prone to outliers compared to the mean, where outliers are distinct values in data. (That is the reason why we use median to impute missing values in categorical features).
- **Mode:** Our last measure of central tendency, the mode is defined as the most frequent value. There may be more than one mode in cases where multiple values are equally frequent.

For example, a die was rolled 10 times, and we got the following numbers: 4, 5, 4, 3, 4, 2, 1, 1, 2, and 1.

The mean is calculated by summing all the events and dividing them by the number of observations: $(4+5+4+3+4+2+1+1+2+1)/10=2.7$.

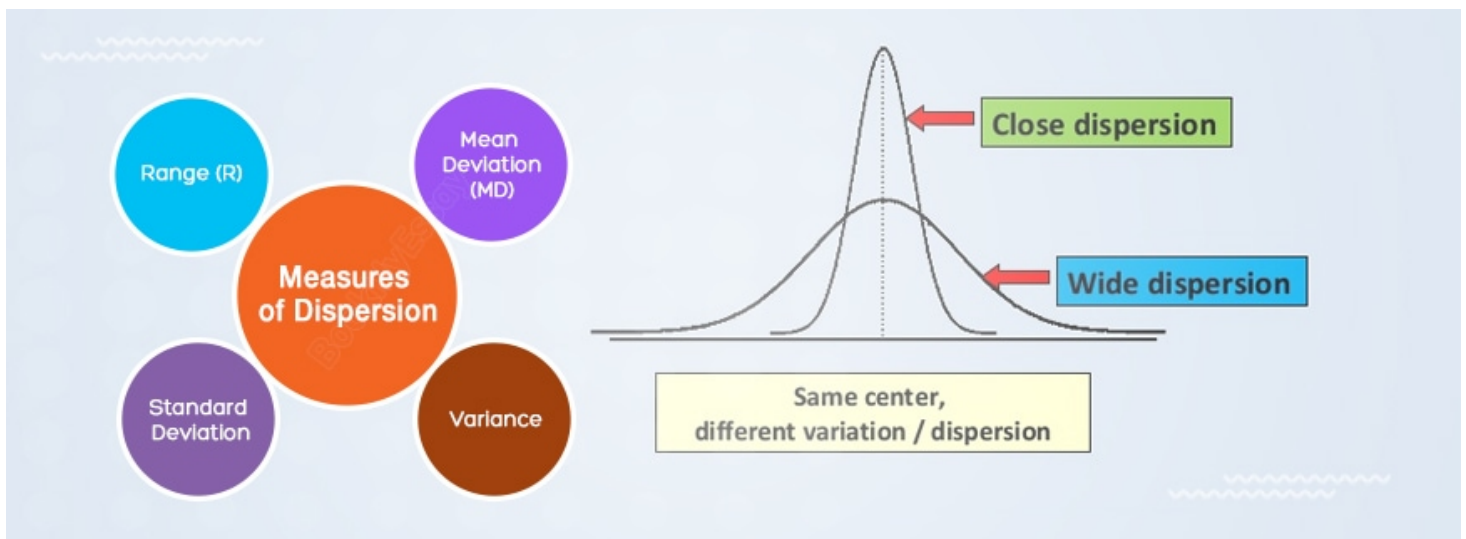
Chapter 1: Importance of Data Visualization

To calculate the median, the die rolls have to be ordered according to their values. The ordered values are as follows: 1, 1, 1, 2, 2, 3, 4, 4, 4, 5. Since we have an even number of die rolls, we need to take the average of the two middle values. The average of the two middle values is $(2+3)/2=2.5$.

The modes are 1 and 4 since they are the two most frequent events.

Measures of Dispersion

Dispersion, also called **variability**, is the extent to which a probability distribution is stretched or squeezed. We can also state dispersion as how the data is distributed across the axis, we use distribution graphs for this (normally histogram, kde).



The different measures of dispersion are as follows:

- **Variance:** The variance is the expected value of the squared deviation from the mean. It describes how far a set of numbers is spread out from their mean.
- **Standard deviation:** This is the square root of the variance.
- **Range:** This is the difference between the largest and smallest values in a dataset.
- **Interquartile range:** Also called the **midspread** or **middle 50%**, this is the difference between the 75th and 25th percentiles, or between the upper and lower quartiles.

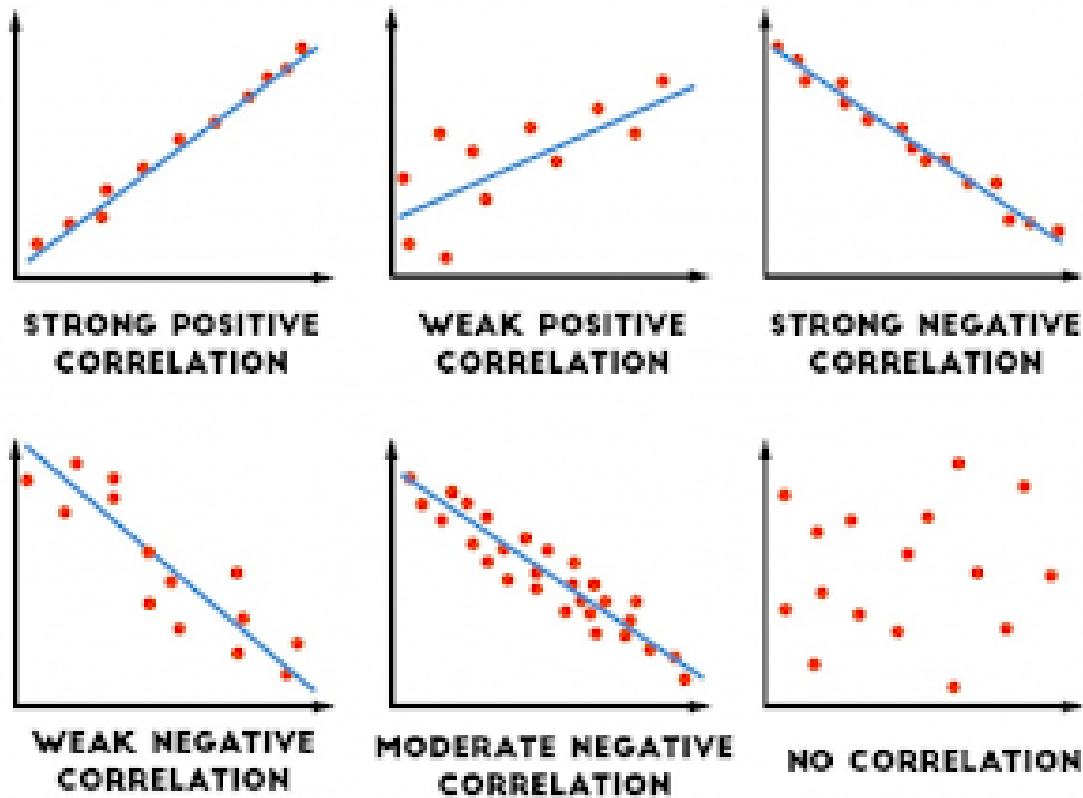
Correlation:

The measures we have discussed so far only considered single variables. In contrast, **correlation** describes the statistical relationship between two variables:

- In a positive correlation, both variables move in the same direction.
- In a negative correlation, the variables move in opposite directions.

Chapter 1: Importance of Data Visualization

- In zero correlation, the variables are not related.



Note

One thing you should be aware of is that correlation does not imply causation. Correlation describes the relationship between two or more variables, while causation describes how one event is caused by another. For example, ice cream sales are correlated with the number of drowning deaths. But that doesn't mean that ice cream consumption causes drowning. There is a third variable, namely temperature, that's responsible for this correlation. Higher temperature causes increasing ice cream sales and more people engaging in swimming, which eventually results in drowning.

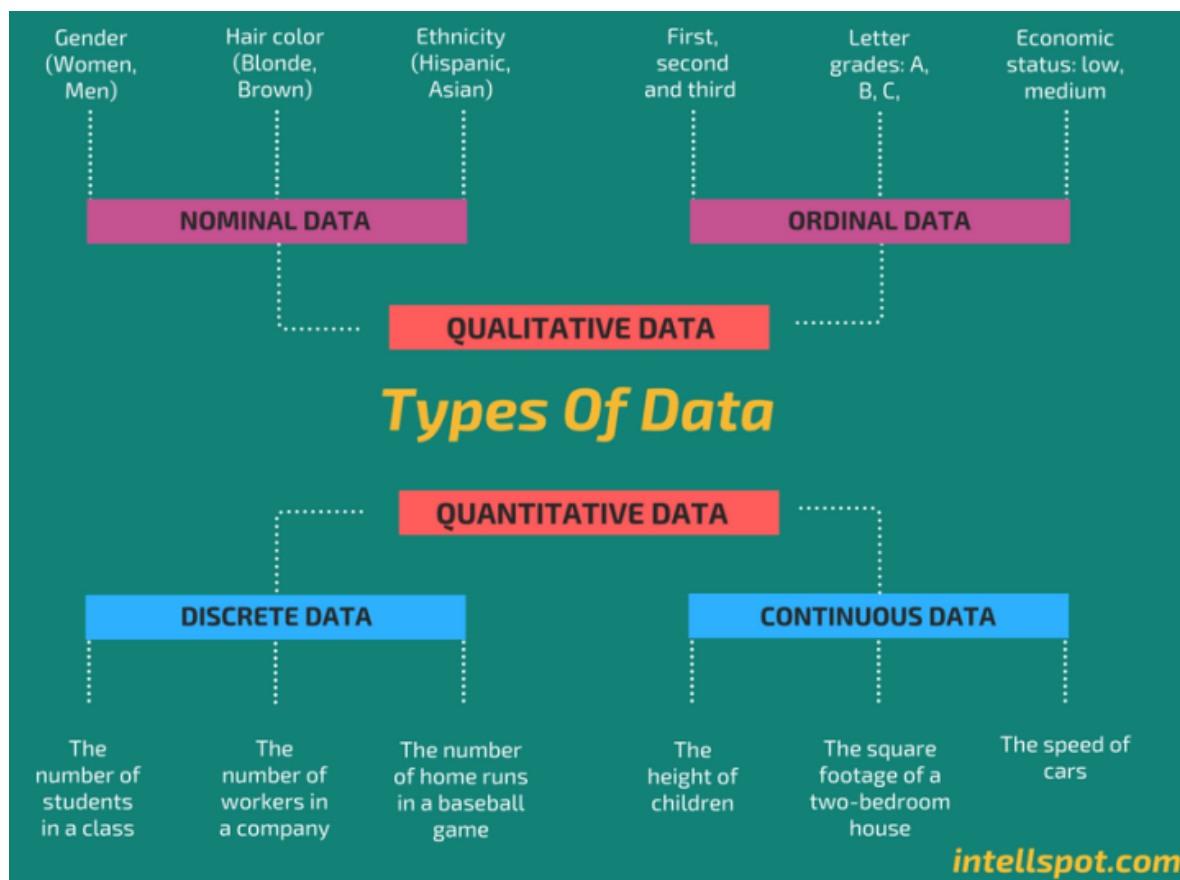
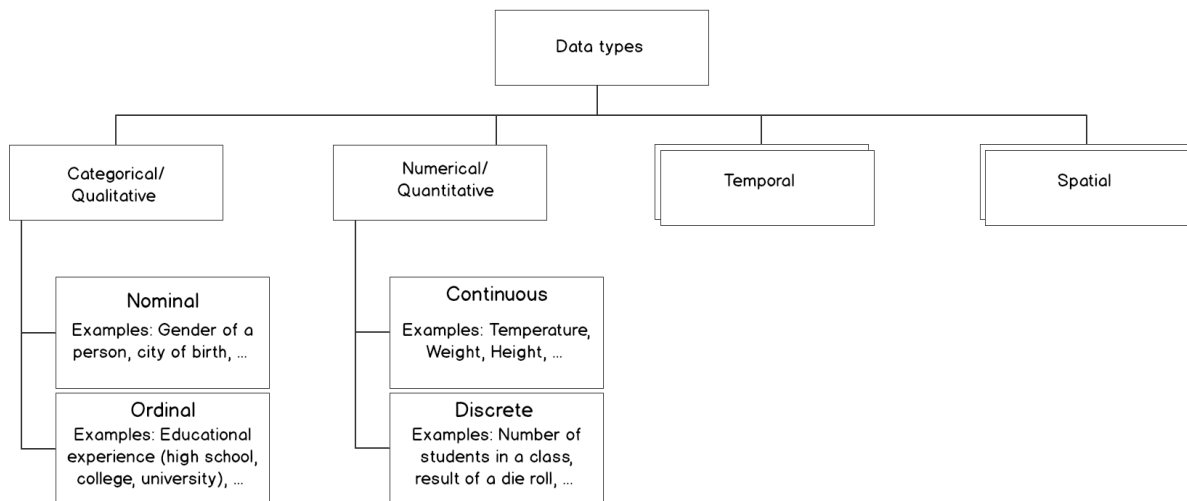
Types of Data

It is important to understand what kind of data you are dealing with so that you can select both the right statistical measure and the right visualization. We categorize data as categorical/qualitative and numerical/quantitative. Categorical data describes characteristics, for example, the colour of an object or a person's gender. We can further divide categorical data into nominal and ordinal data. In contrast to nominal data, ordinal data has an order.

Chapter 1: Importance of Data Visualization

Numerical data can be divided into discrete and continuous data. We speak of discrete data if the data can only have certain values, whereas continuous data can take any value (sometimes limited to a range).

Another aspect to consider is whether the data has a temporal domain – in other words, is it bound to time or does it change over time? If the data is bound to a location, it might be interesting to show the spatial relationship, so you should keep that in mind as well:



Chapter 1: Importance of Data Visualization

The above definitions are what statistics define. According to me there are 11 different types of data that a machine learning practitioner deals with, they are:

1. Useless: Data which is useless for the model. Ex. Features in our dataset which are constant for all samples, features in our dataset which has unique value for every sample (Id etc.).
2. Nominal
3. Binary: Features in our dataset which just has two unique values for all samples. Ex. Gender feature (M/F), features having just yes or no.
4. Ordinal
5. Count: Features in our dataset that reflects the count of any dependent variable.
6. Time
7. Interval
8. Image
9. Video
10. Audio
11. Text

The following table gives an overview of which measure of central tendency is best suited to a particular type of data:

Data type	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Numerical	Mean/Median

Chapter 2: Everything about plots

Introduction

In this chapter, we will focus on various visualizations and identify which visualization is best for showing certain information for a given dataset. We will describe every visualization in detail and give practical examples, such as comparing different stocks over time or comparing the ratings for different movies. Starting with comparison plots, which are great for comparing multiple variables over time, we will look at their types (such as line charts and bar charts).

We will then move onto relation plots, which are handy for showing relationships among variables. We will cover scatter plots for showing the relationship between two variables, bubble plots for three variables, correlograms for variable pairs, and finally, heatmaps for visualizing multivariate data.

The chapter will further explain composition plots (used to visualize variables that are part of a whole), as well as pie charts, stacked bar charts and stacked area charts. To give you a deeper insight into the distribution of variables, we will discuss distribution plots, describing histograms, density plots, box plots, and violin plots.

Comparison Plots

Comparison plots include charts that are ideal for comparing multiple variables or variables over time.



Chapter 2: Everything about plots

Line Chart

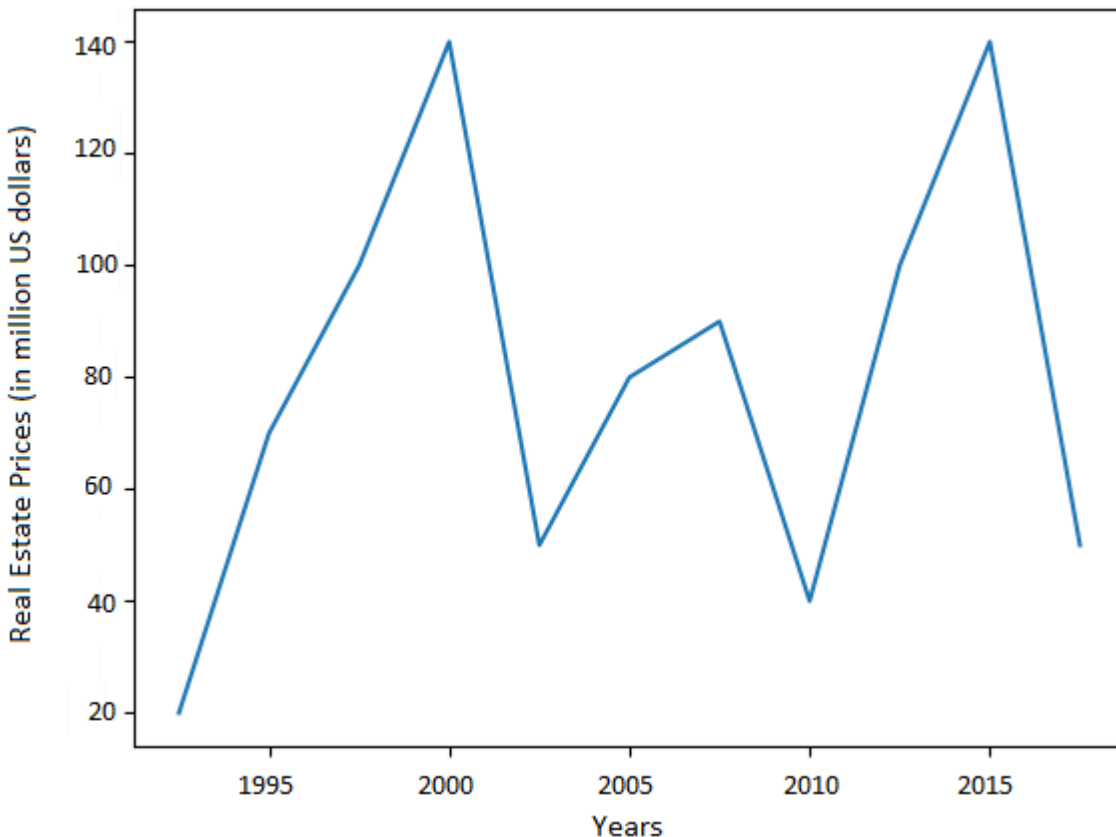
Line charts are used to display quantitative values over a continuous time period and show information as a series. A line chart is ideal for a time series that is connected by straight-line segments.

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.

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:



Design Practices

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

Chapter 2: Everything about plots

Note

For plots with multiple variables, a legend should be given to describe each variable.

Bar Chart

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

Use

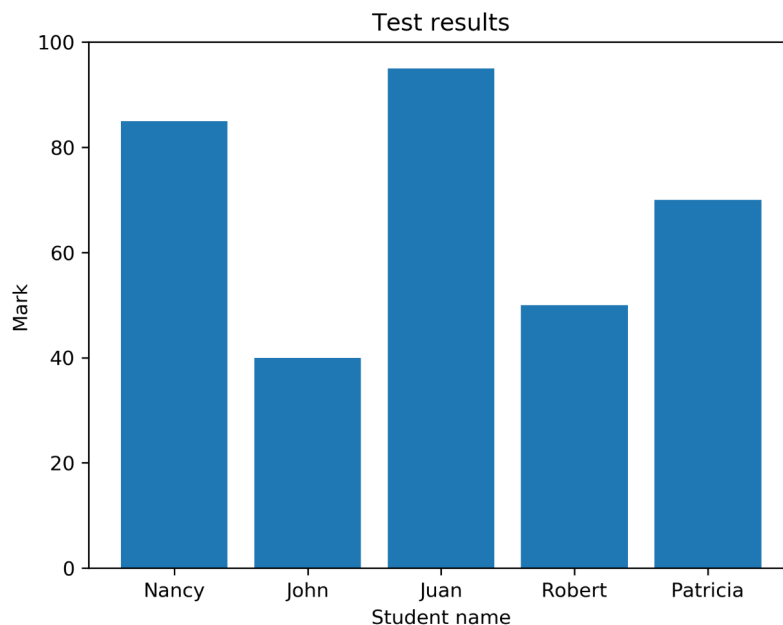
While they are both used to compare numerical values across categories, vertical bar charts are sometimes used to show a single variable over time.

Don'ts of Bar Charts

- Don't confuse vertical bar charts with histograms. Bar charts compare different variables or categories, while histograms show the distribution for a single variable. Histograms will be discussed later in this chapter.
- Another common mistake is to use bar charts to show central tendencies among groups or categories. Use box plots or violin plots to show statistical measures or distributions in these cases, which will also be discussed later in this chapter.

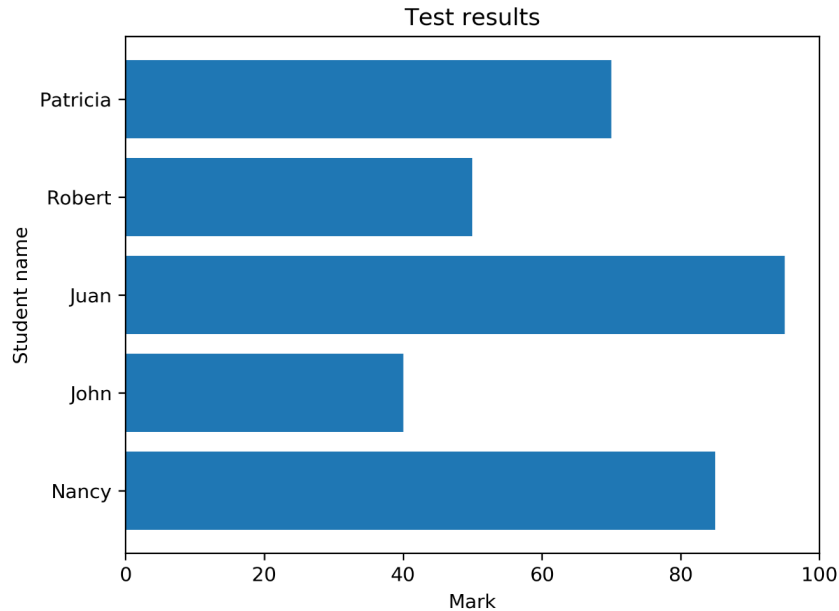
Examples

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

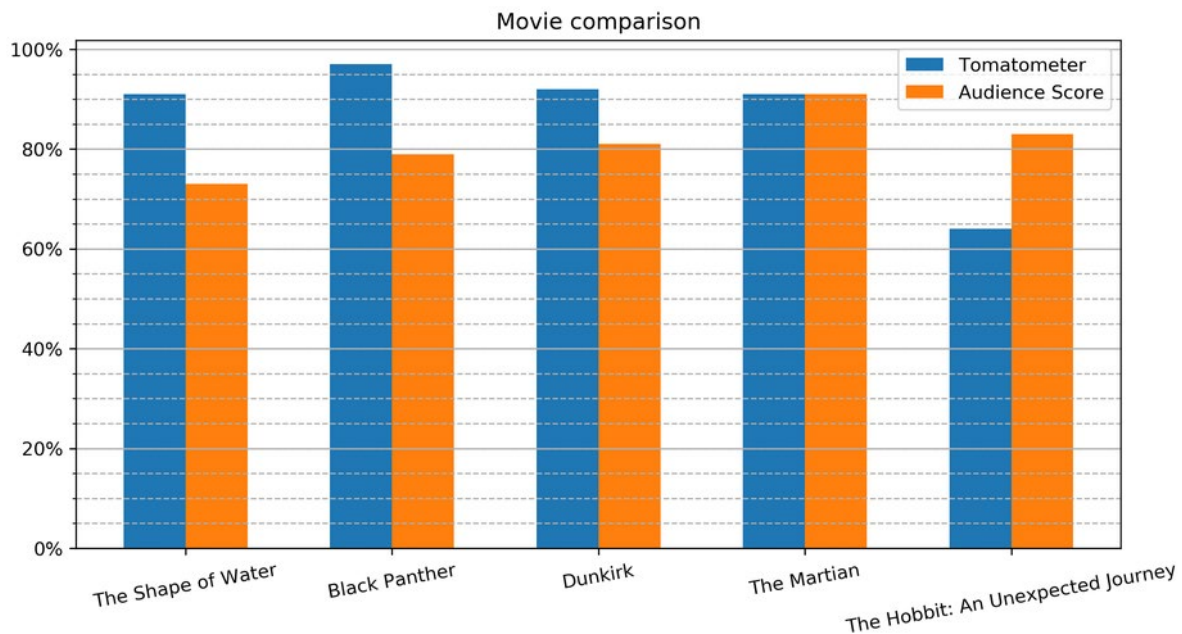


Chapter 2: Everything about plots

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



The following diagram compares movie ratings, giving two different scores. The Tomatometer is the percentage of approved critics who have given a positive review for the movie. The Audience Score is the percentage of users who have given a score of 3.5 or higher out of 5. As we can see, **The Martian** is the only movie with both a high Tomatometer and Audience Score. **The Hobbit: An Unexpected Journey** has a relatively high Audience Score compared to the Tomatometer score, which might be due to a huge fan base:



Chapter 2: Everything about plots

Design Practices

- The axis corresponding to the numerical variable should start at zero. Starting with another value might be misleading, as it makes a small value difference look like a big one.
- Use horizontal labels—that is, as long as the number of bars is small, and the chart doesn't look too cluttered.
- The labels can be rotated to different angles if there isn't enough space to present them horizontally. You can see this on the labels of the x-axis of the preceding diagram.

(`plt.xticks(rotation = angle to be rotated)`)

Relation Plots

Relation plots are perfectly suited to showing relationships among variables.

Scatter Plot

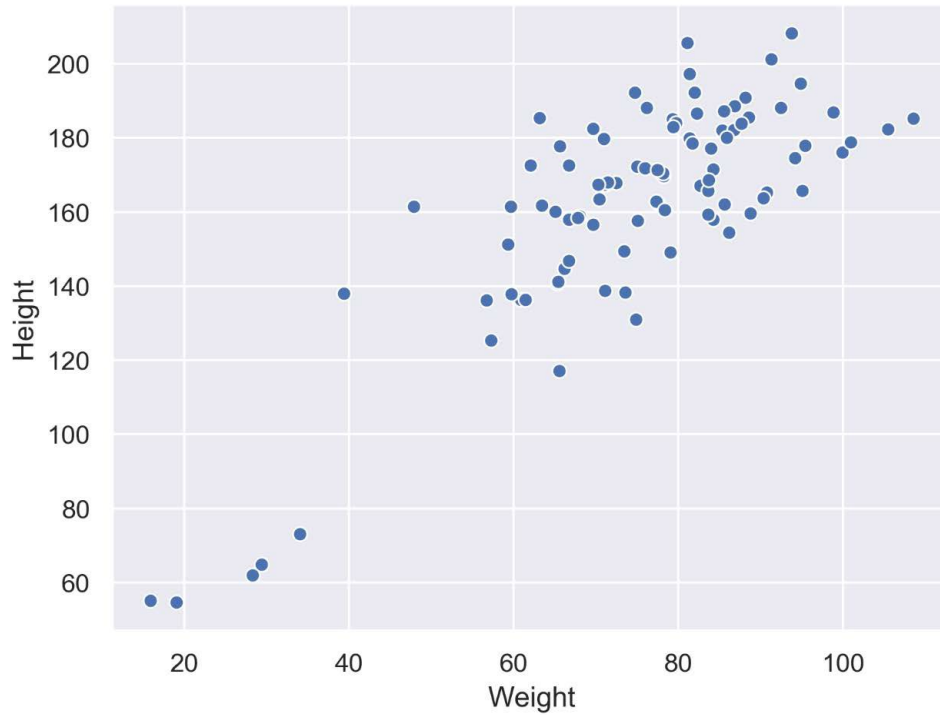
Scatter plots show data points for two numerical variables, displaying a variable on both axes.

- You can detect whether a correlation (relationship) exists between two variables.
- They allow you 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.

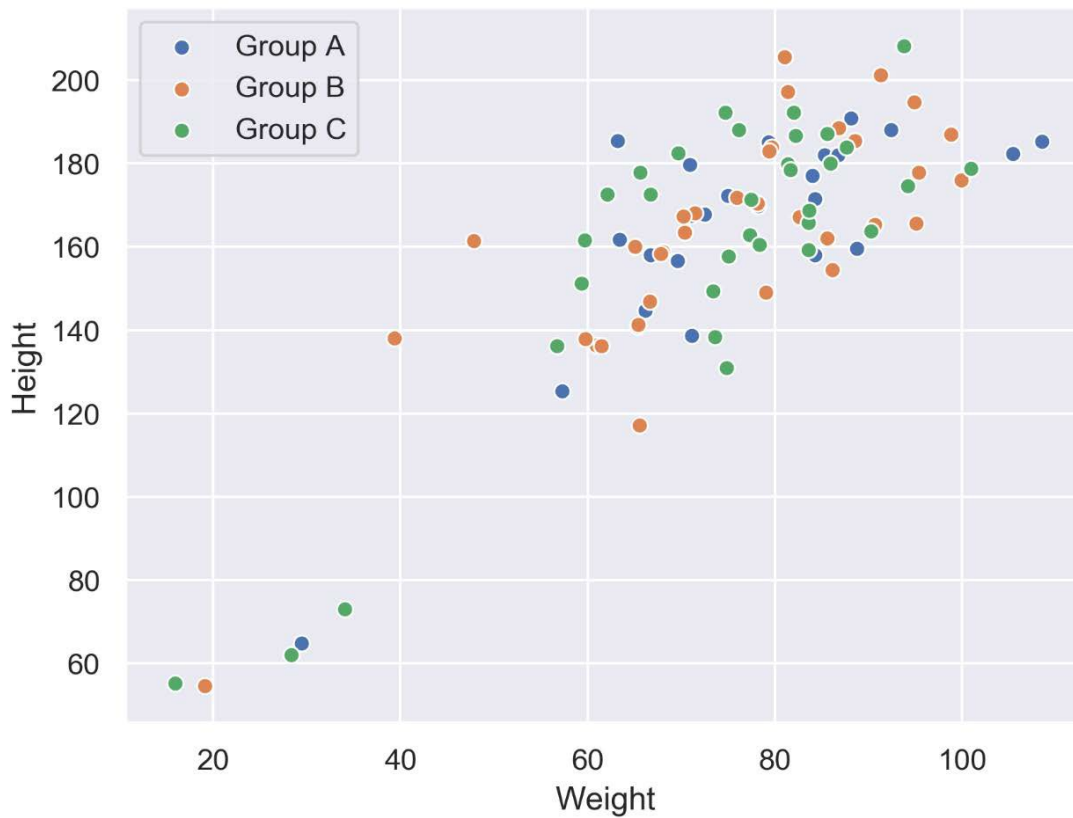
Examples

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

Chapter 2: Everything about plots

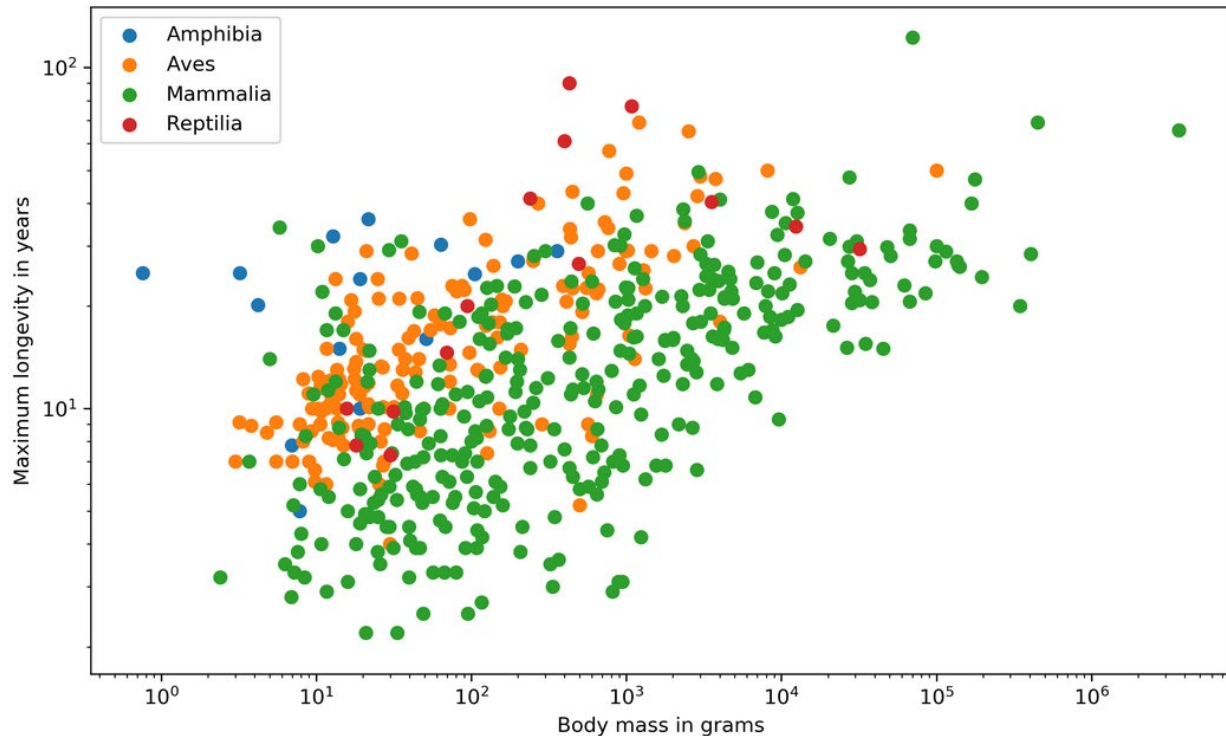


The following diagram shows the same data as in the previous plot but differentiates between groups. In this case, we have different groups: **A**, **B**, and **C**:



Chapter 2: Everything about plots

The following diagram shows the correlation between body mass and the maximum longevity for various animals grouped by their classes. There is a positive correlation between body mass and maximum longevity:



Design Practices

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

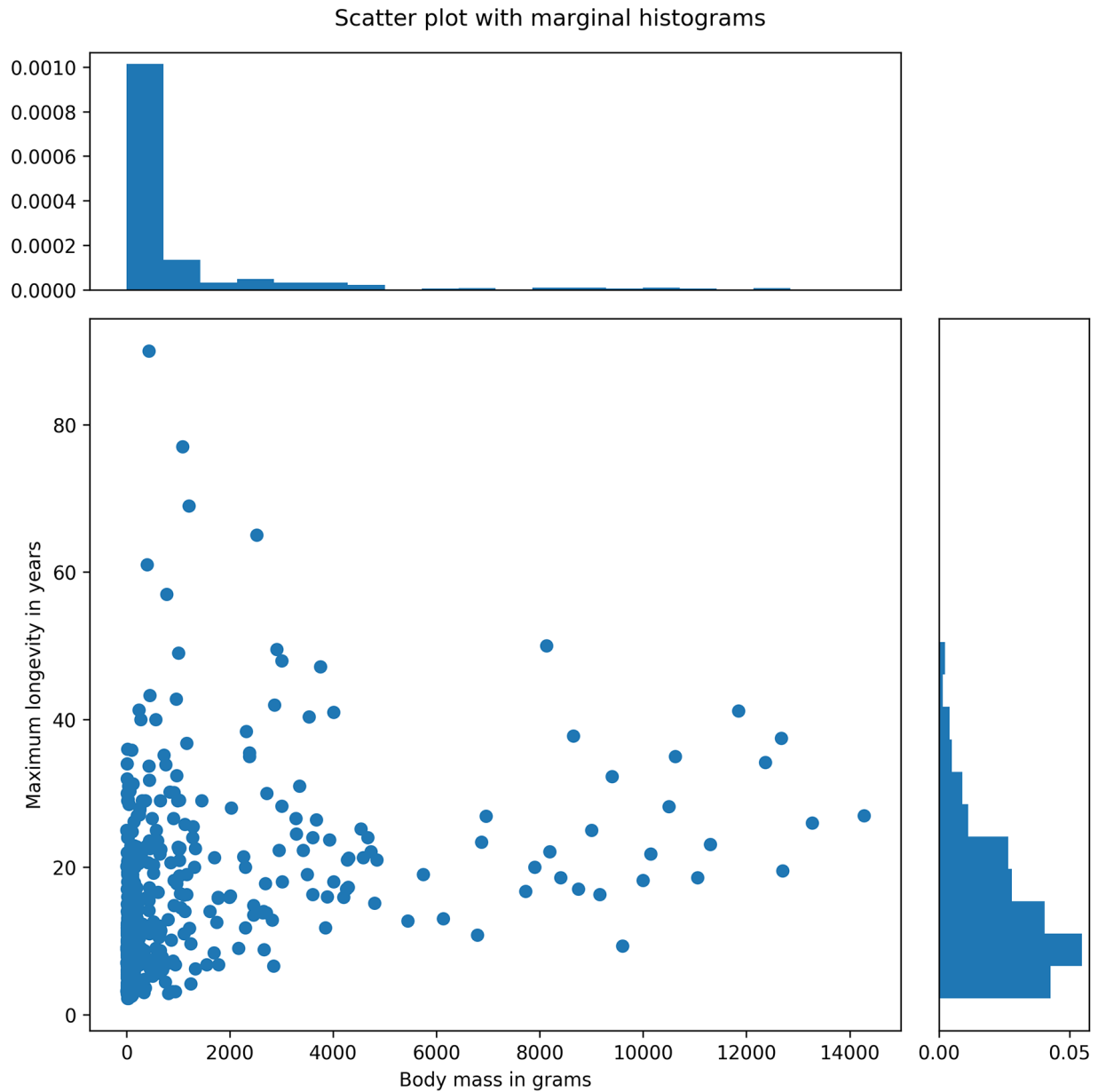
Variants: Scatter Plots with Marginal Histograms

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. Distribution graphs will be discussed later in the chapter.

Examples

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:

Chapter 2: Everything about plots



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.

Use

Bubble plots help to show a correlation between three variables.

Example

Chapter 2: Everything about plots

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:



Design Practices

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

Correlogram

A **correlogram** is a combination of scatter plots and histograms. Histograms will be discussed in detail later in this chapter. A correlogram or correlation matrix visualizes the relationship between each pair of numerical variables using a scatter plot.

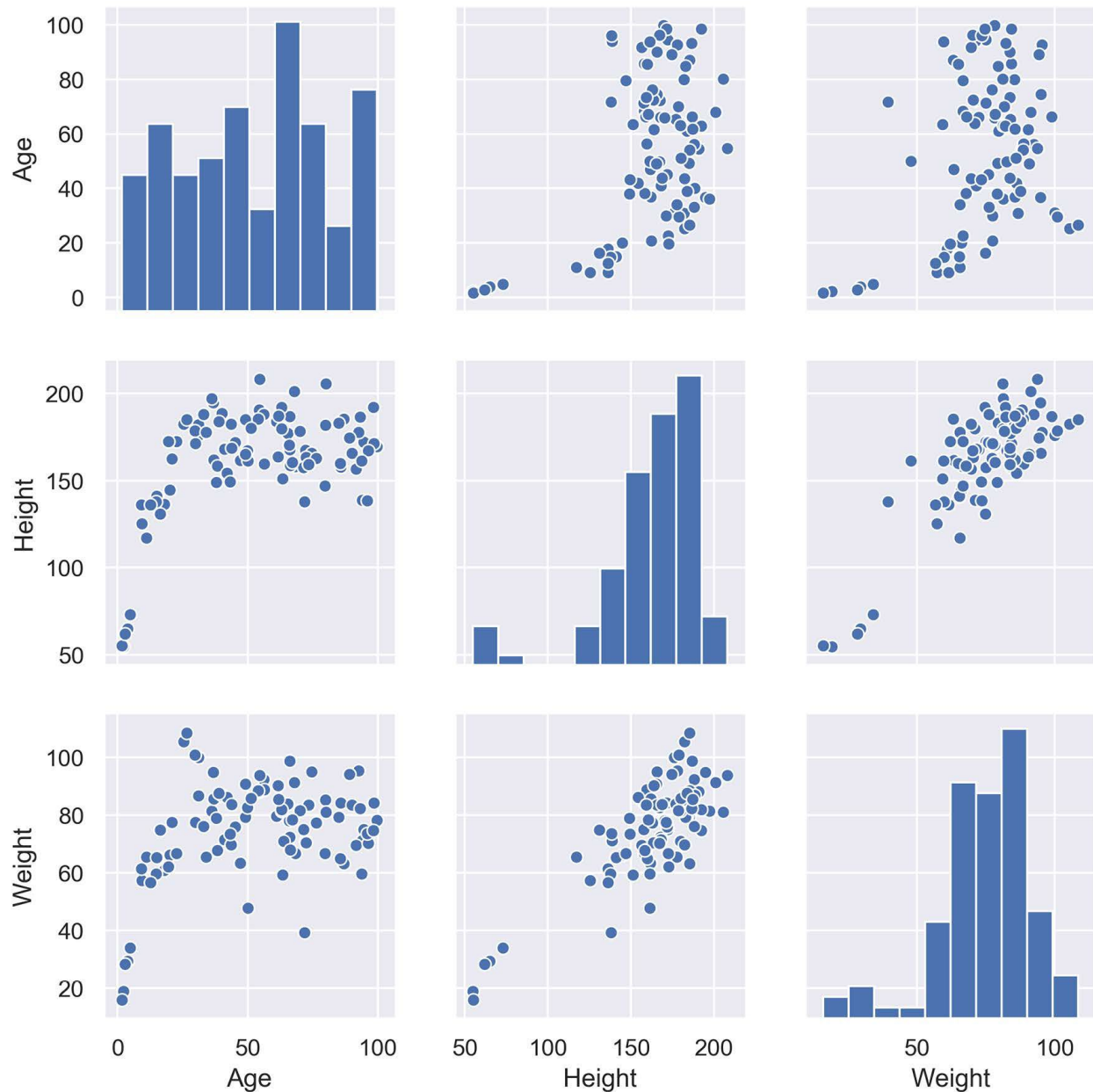
The diagonals of the correlation matrix represent the distribution of each variable in the form of a histogram. You can also plot the relationship between multiple groups or categories using

Chapter 2: Everything about plots

different colors. A correlogram is a great chart for exploratory data analysis to get a feel for your data, especially the correlation between variable pairs.

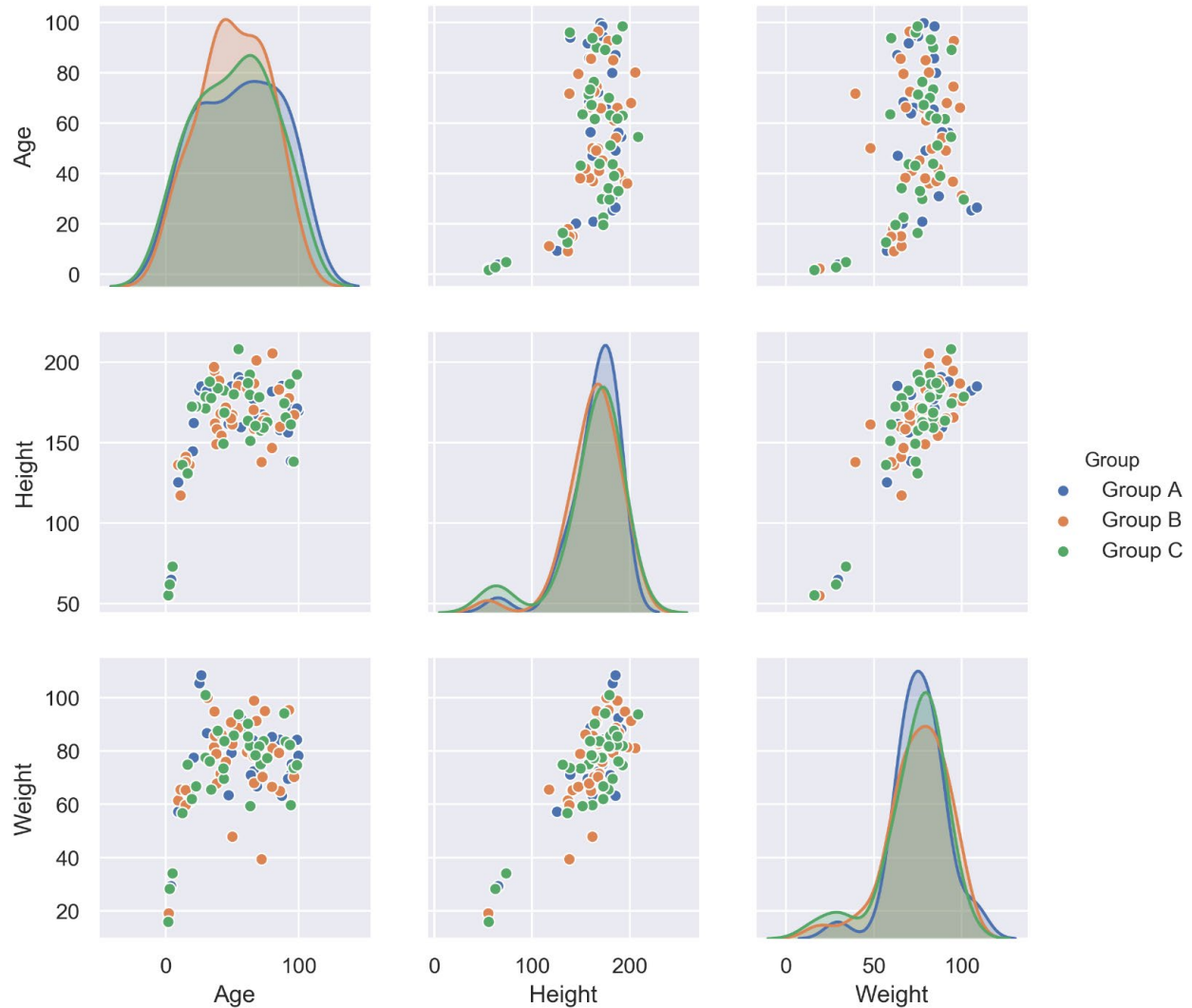
Examples

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:



The following diagram shows the correlogram with data samples separated by color into different groups, here instead of histograms, we'll be using area under graph to visualize the distributions (area under graph/density plots will be discussed later in the chapter):

Chapter 2: Everything about plots



Design Practices

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

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), where categorical variables are placed in the rows and columns and a numerical or categorical variable is represented as colors or color saturation.

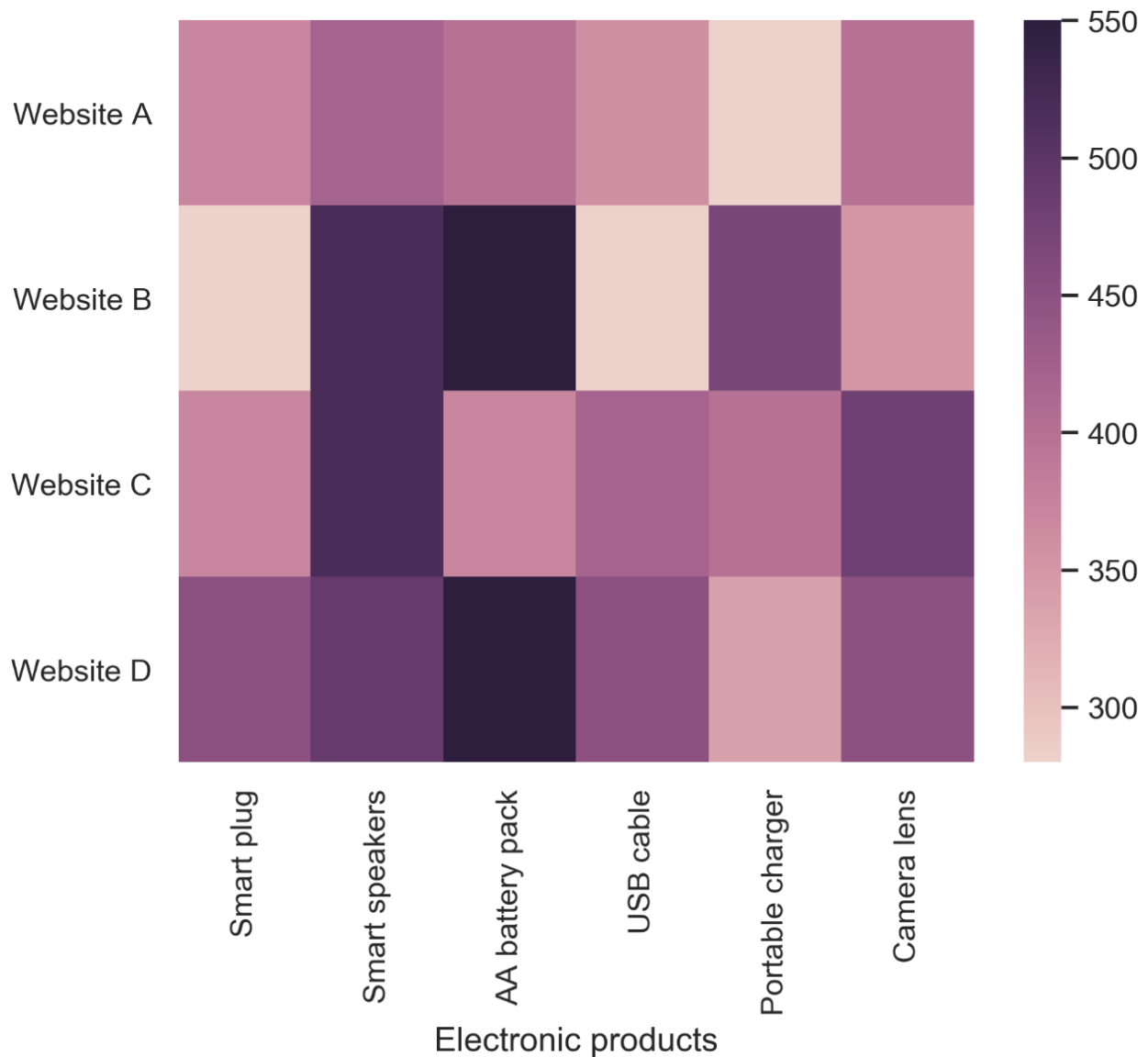
Chapter 2: Everything about plots

Use

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

Examples

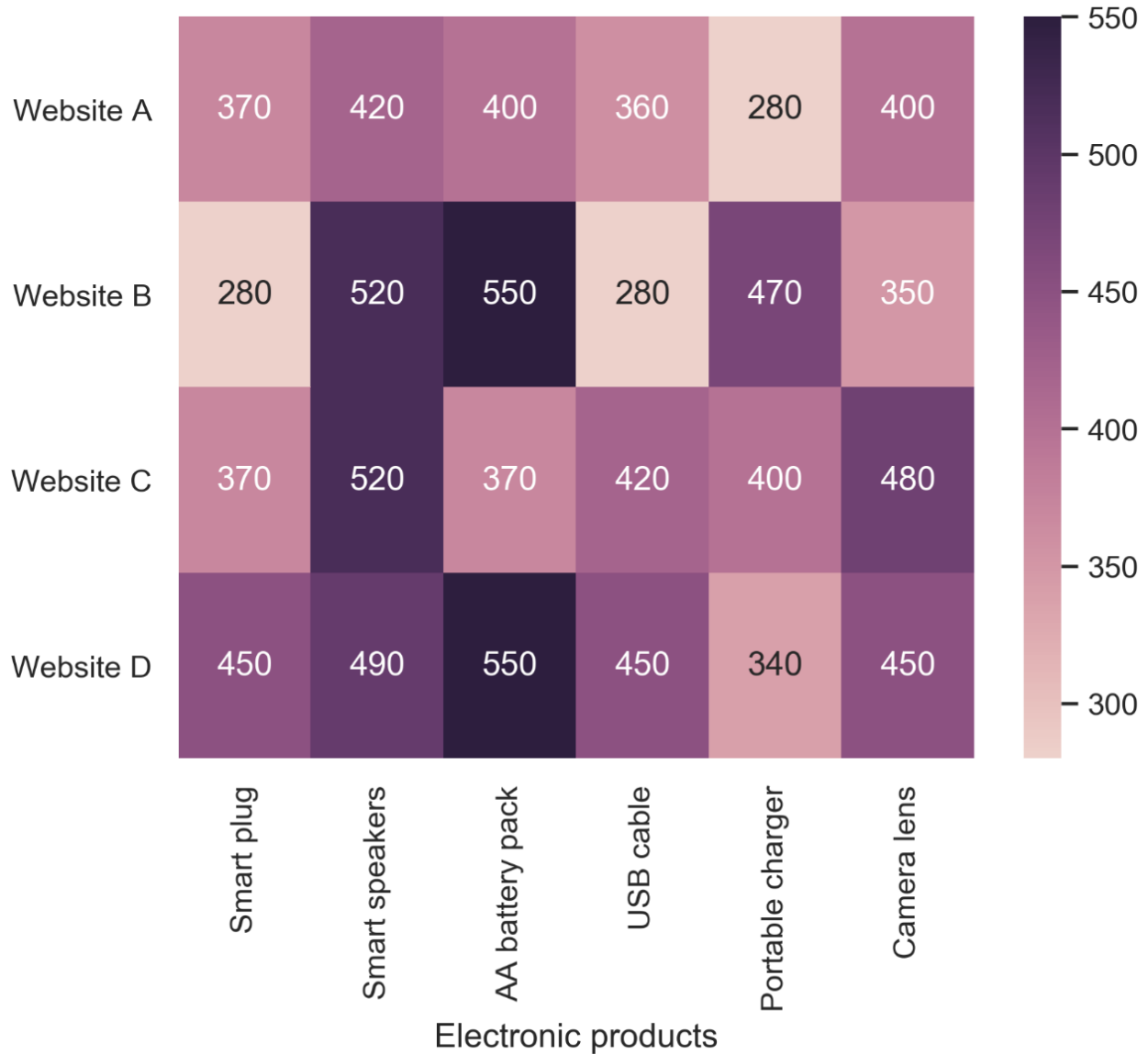
The following diagram shows a heatmap for the most popular products on the electronics category page across various e-commerce websites, where the color shows the number of units sold. In the following diagram, we can analyze that the darker colors represent more units sold, as shown in the key:



Chapter 2: Everything about plots

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:



Design Practice

- Select colors and contrasts that will be easily visible to individuals with vision problems so that your plots are more inclusive.

Chapter 2: Everything about plots

Composition Plots



Composition plots are ideal if you think about something as a part of a whole.

Since data scientists rarely use composition plots, so I'll be skipping pie charts and venn diagrams and only discuss about stacked bar charts and stacked area charts.

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. You can either compare total amounts across each bar or show a percentage of each group. 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.

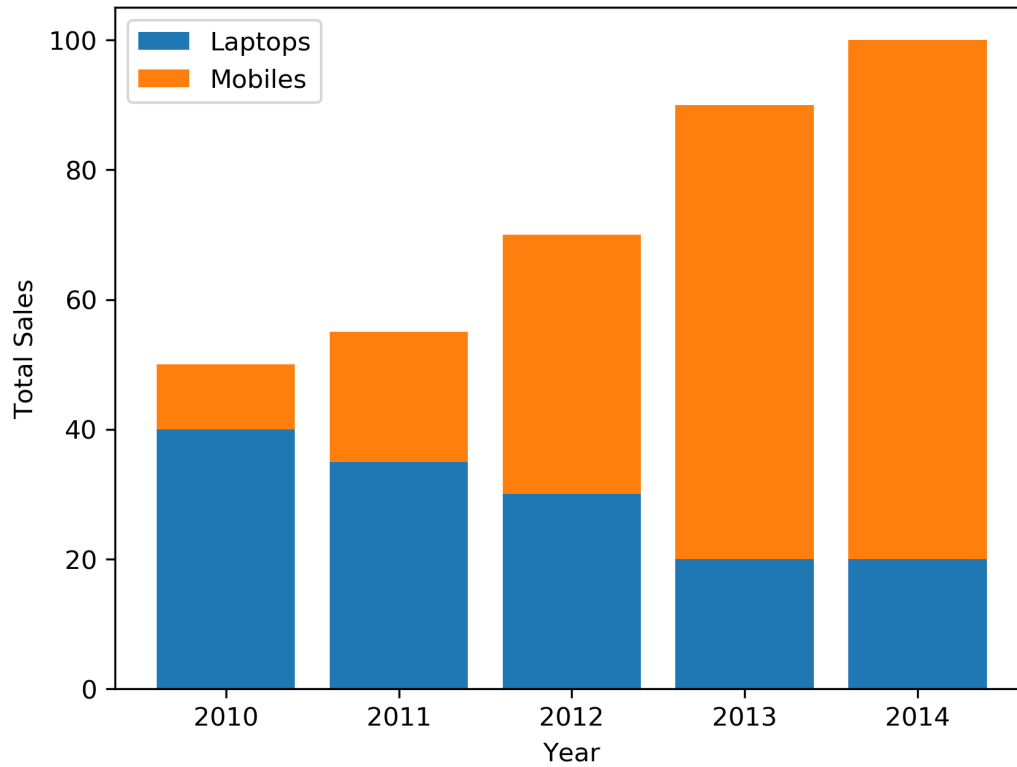
Use

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

Examples

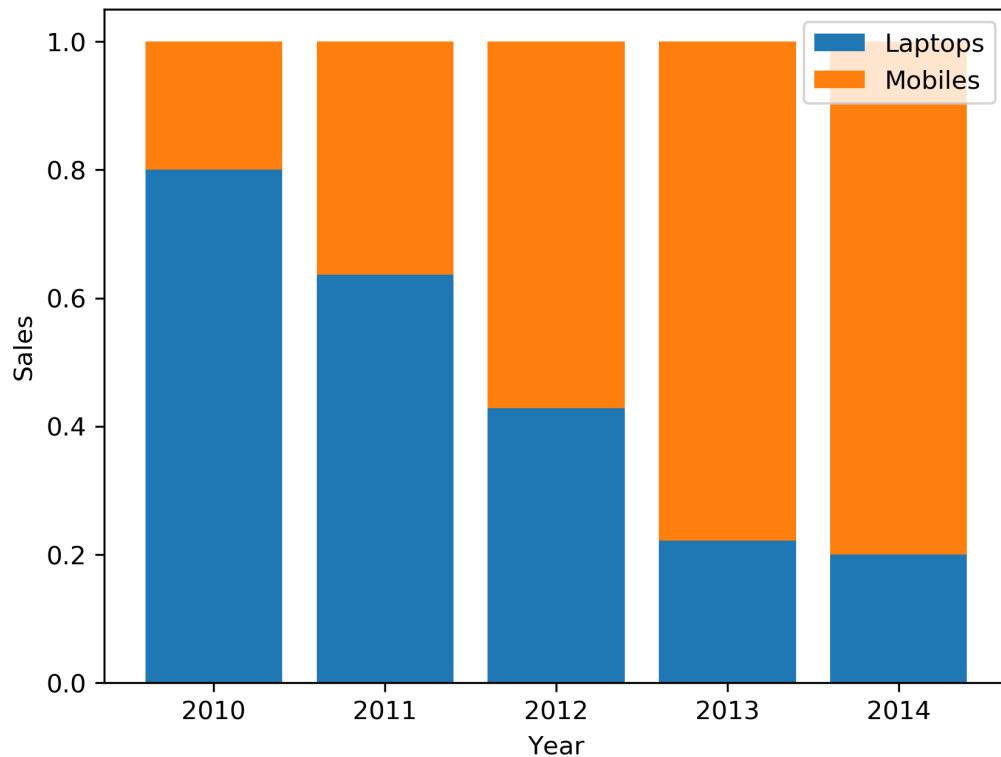
The following diagram shows a generic stacked bar chart with five groups:

Chapter 2: Everything about plots



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:



Chapter 2: Everything about plots

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.

Stacked Area Chart

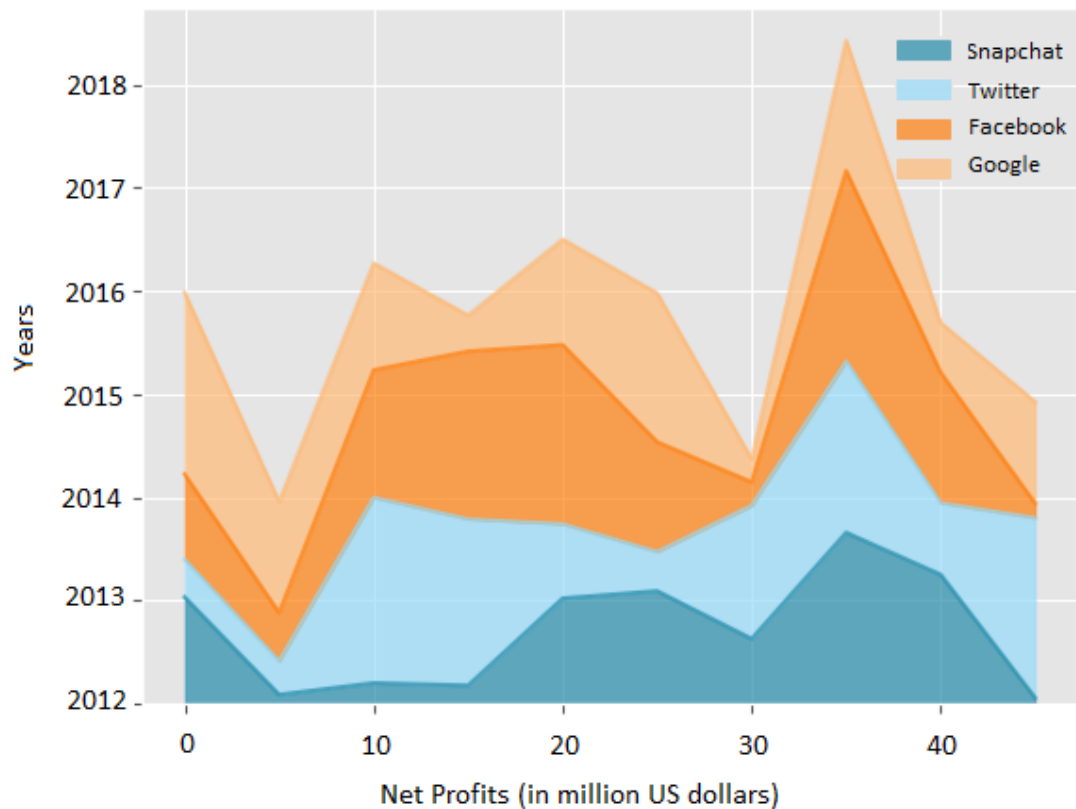
Stacked area charts show trends for part-of-a-whole relations. The values of several groups are illustrated by stacking individual area charts on top of one another. It helps to analyze both individual and overall trend information.

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:



Chapter 2: Everything about plots

Design Practice

- Use transparent colors to improve information visibility. This will help you to analyze the overlapping data and you will also be able to see the grid lines.

Distribution Plots

Distribution plots give a deep insight into how your data is distributed.

This kind of plot is most used by ml practitioner for EDA. Distribution plots give insights into the statistics for each feature. By using this knowledge great features can be generated.

Histogram

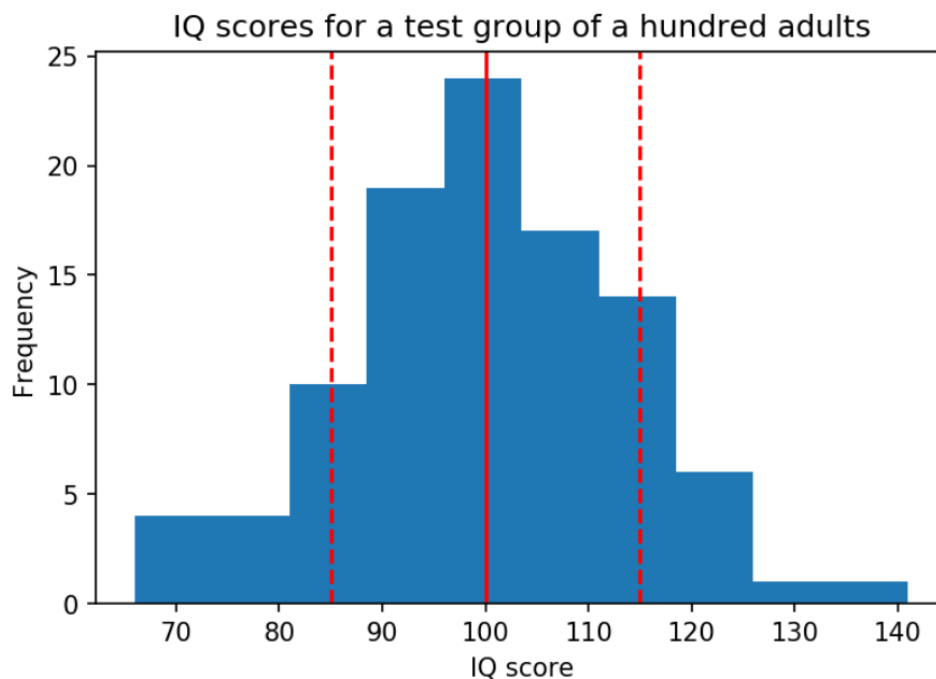
A **histogram** visualizes the distribution of a single numerical variable. Each bar represents the frequency for a certain interval. Histograms help get an estimate of statistical measures. You see where values are concentrated, and you can easily detect outliers. You can either plot a histogram with absolute frequency values or, alternatively, normalize your histogram. If you want to compare distributions of multiple variables, you can use different colors for the bars.

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):



Chapter 2: Everything about plots

Design Practice

- Try different numbers of bins (data intervals), since the shape of the histogram can vary significantly (very important thing to keep in mind, the distribution might be hidden due to wrong selection of bins, usually vary bin size from 5-20. If still not satisfied plot density plots).

Density Plot

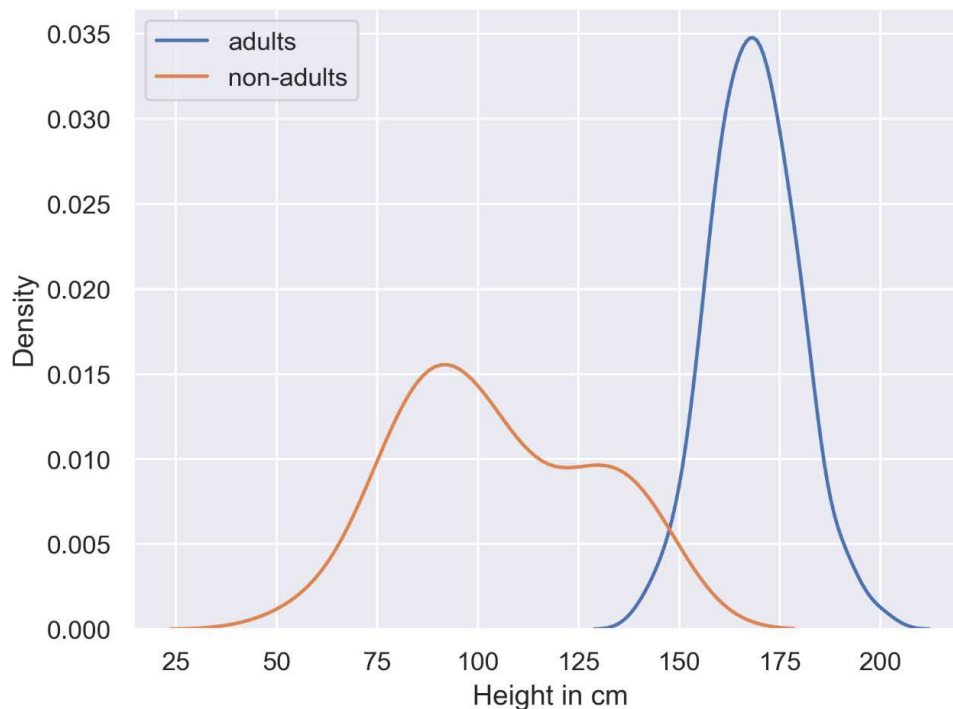
A **density plot** shows the distribution of a numerical variable. It is a variation of a histogram that uses **kernel smoothing**, allowing for smoother distributions. One advantage these have over histograms is that density plots are better at determining the distribution shape since the distribution 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 multi - density plot:



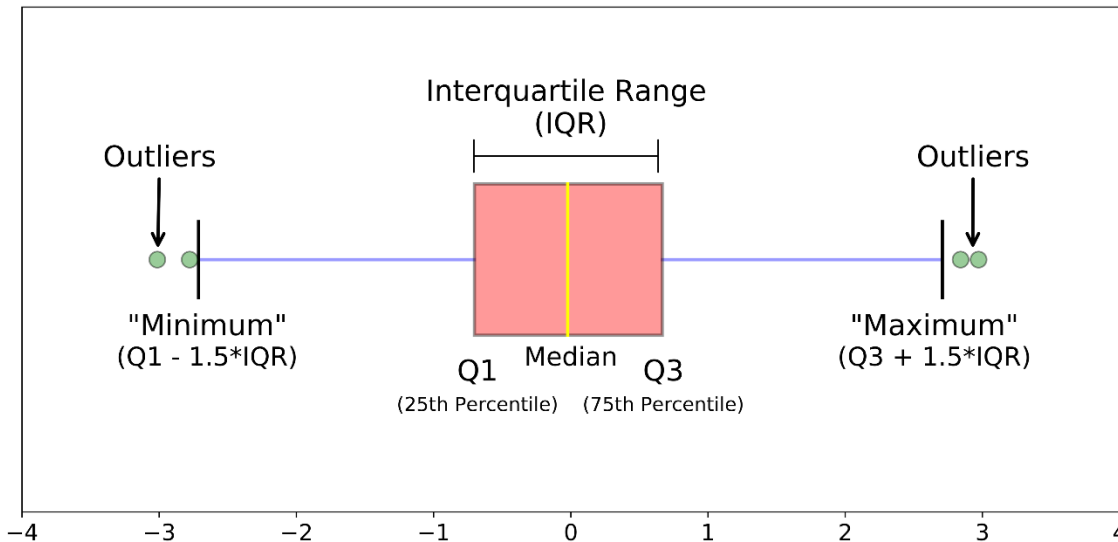
Design Practice

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

Chapter 2: Everything about plots

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.



The image above is a boxplot. A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

To understand box plots and violin plots we must understand few statistical definitions and terminologies.

Boxplots are a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").

Median (Q2/50th Percentile): the middle value of the dataset.

First quartile (Q1/25th Percentile): the middle number between the smallest number (not the "minimum") and the median of the dataset.

Third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the "maximum") of the dataset.

Chapter 2: Everything about plots

Interquartile range (IQR): 25th to the 75th percentile.

Whiskers (shown in blue)

Outliers (shown as green circles)

“Maximum”: $Q3 + 1.5 \cdot IQR$

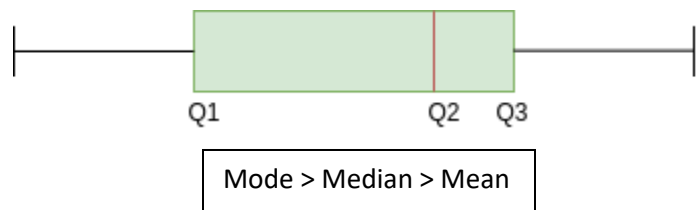
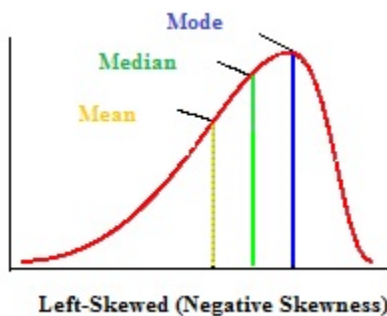
“Minimum”: $Q1 - 1.5 \cdot IQR$

Skewed data

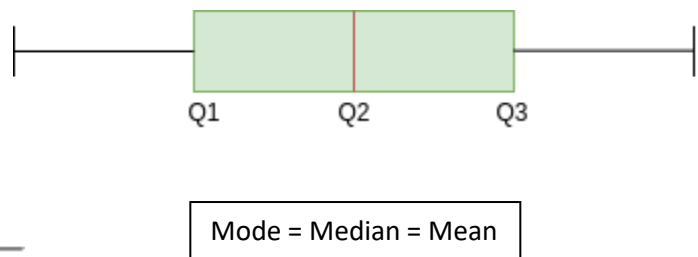
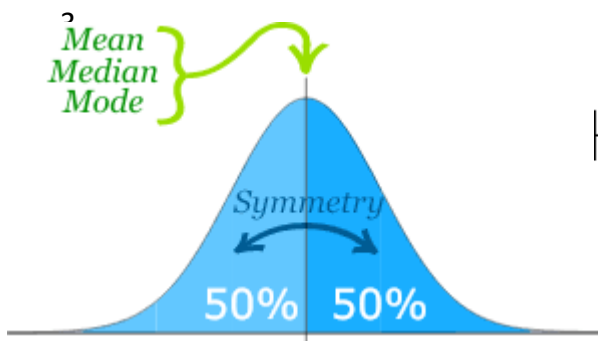
The machine learning model always like symmetrical data (also known as data distributed normally/ Gaussian distribution/ Bell curve), so that it can make predictions easily. But in reality all the data features in a dataset are not symmetrical, they have a long tail on one side or the other and they are known as skewed features. To get a better model we must always transform our data and make skewed features as less skewed as possible.

When talking about symmetry of data, it can be divided into 3 types. They are:

1. **Negative skew:** The long tail is on left side of the peak. People sometimes say it is “skewed to the left”.

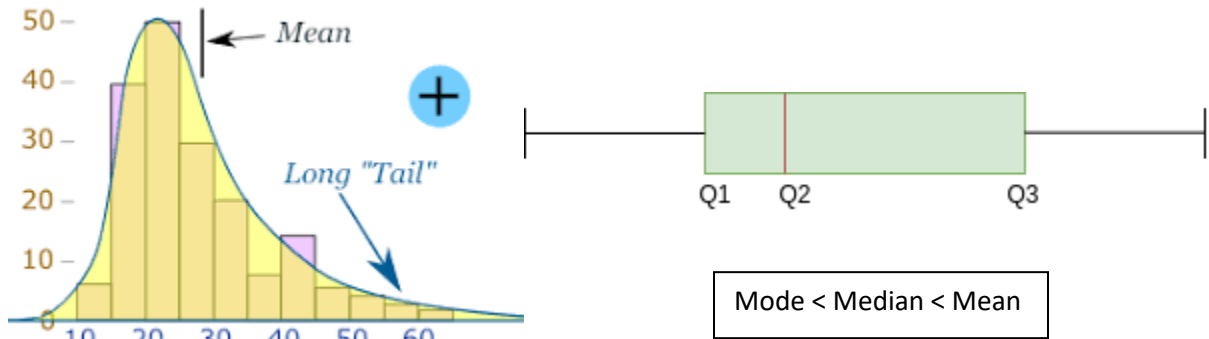


2. **Normal distribution (no skew):** A normal distribution is not skewed. It is perfectly symmetrical.

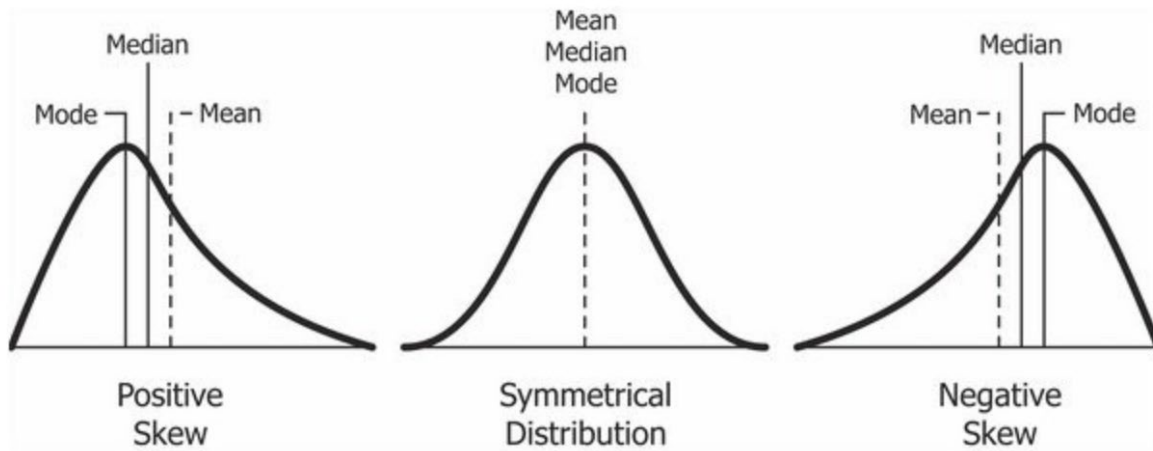


Chapter 2: Everything about plots

3. **Positive skew:** The long tail is on the positive side of the peak, and people say it is "skewed to the right".

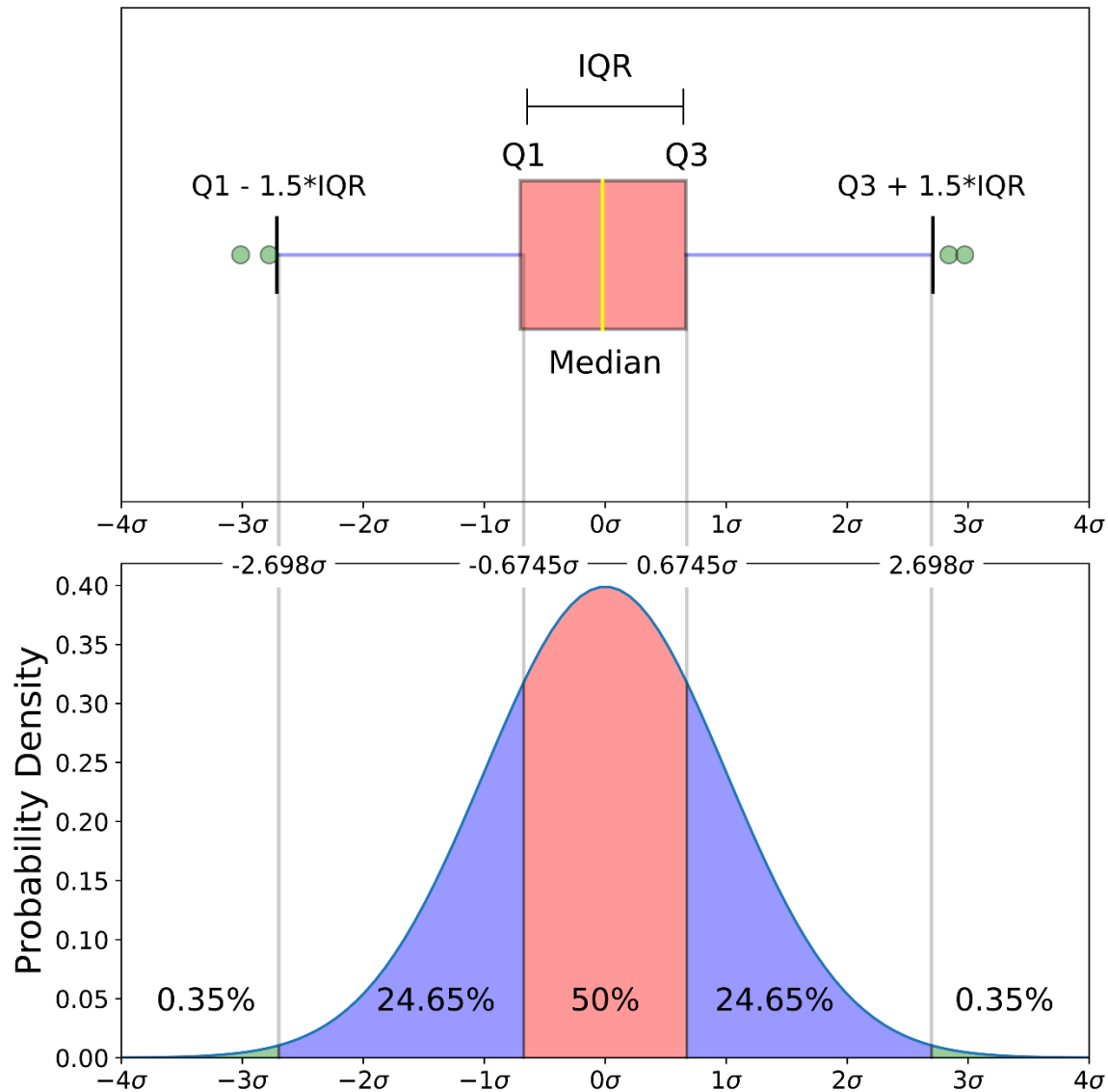


To summarize skewed data:



Chapter 2: Everything about plots

Boxplot on Normal Distribution:



The image above is a comparison of a boxplot of a nearly normal distribution and the probability density function (pdf) for a normal distribution. The reason why I am showing you this image is that looking at a statistical distribution is more commonplace than looking at a box plot. In other words, it might help you understand a boxplot.

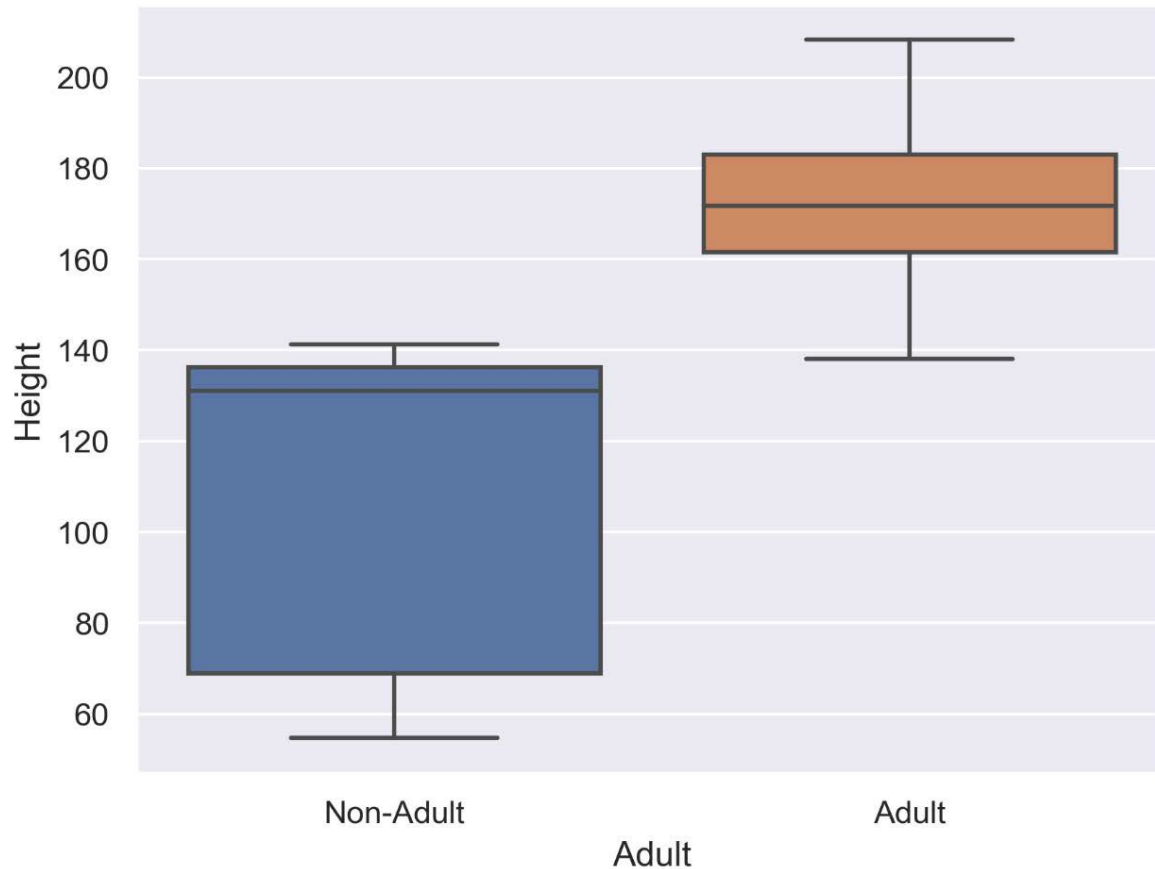
Use

Compare statistical measures for multiple variables or groups.

Chapter 2: Everything about plots

Examples

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:



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.

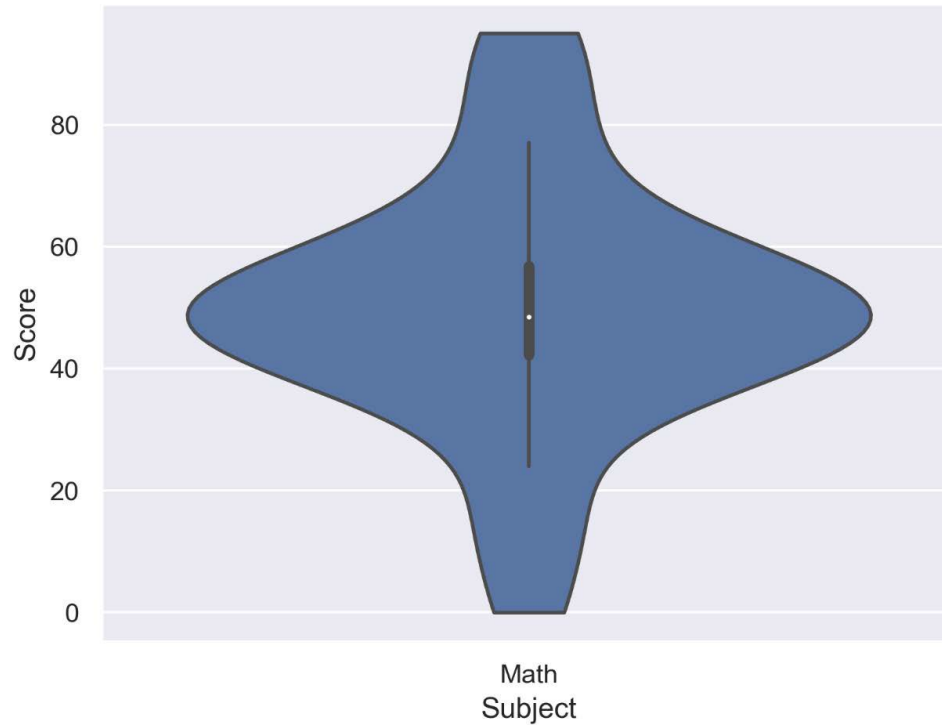
Use

Compare statistical measures and density for multiple variables or groups.

Examples

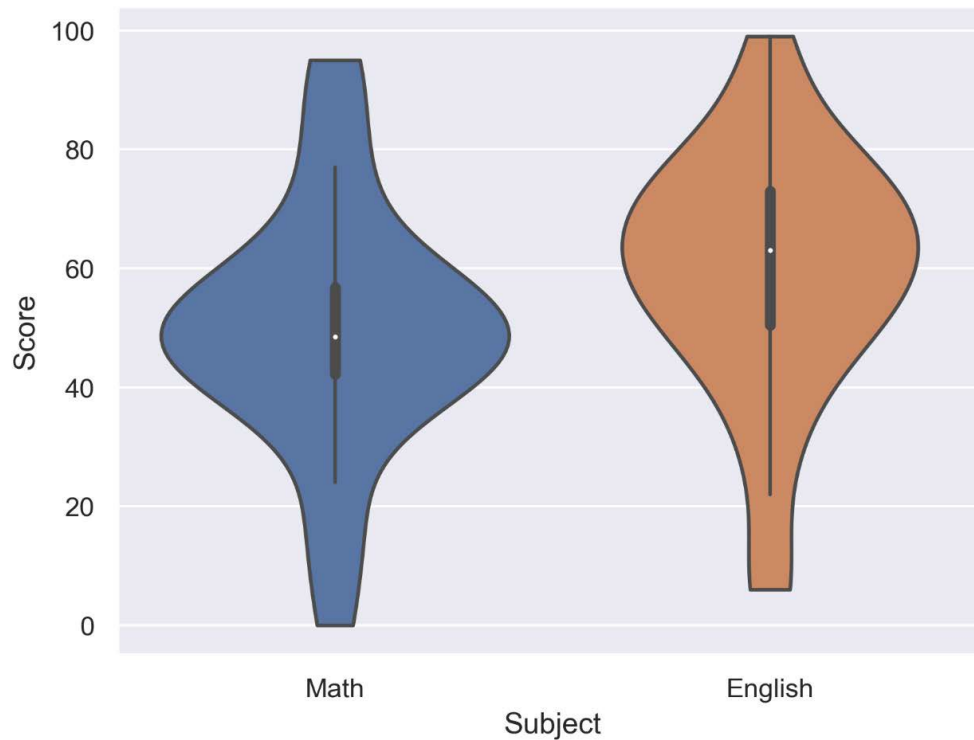
The following diagram shows a violin plot for a single variable and shows how students have performed in **Math**:

Chapter 2: Everything about plots



From the preceding diagram, we can analyze that most of the students have scored around 40-60 in the **Math** test.

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



Chapter 2: Everything about plots

From the preceding diagram, we can say that on average, the students have scored more in **English** than in **Math**, but the highest score was secured in **Math**.

Design Practice

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

Summary

This chapter covered the most important visualizations, categorized into comparison, relation, composition and distribution

Line charts are great for comparing something over time.

Bar charts are great for comparing different items.

A scatter plot visualizes the correlation between 2 variables for one or multiple groups.

Bubble plots can be used to show relationship between 3 variables. The additional 3rd variable is represented by the dot size.

Heatmaps are great for revealing patterns and correlations between two qualitative variables.

A correlogram is a perfect visualization for showing the correlation among multiple variables.

To show proportions and percentages for groups, we can use either stacked bar charts or stacked area charts.

For a single variable histogram is effective for distribution plots.

For multiple variables, you can either use a boxplot or a violin plot. The violin plot visualizes densities of your variables whereas the box plot just visualizes the median, the IQ range and the range for each variable.