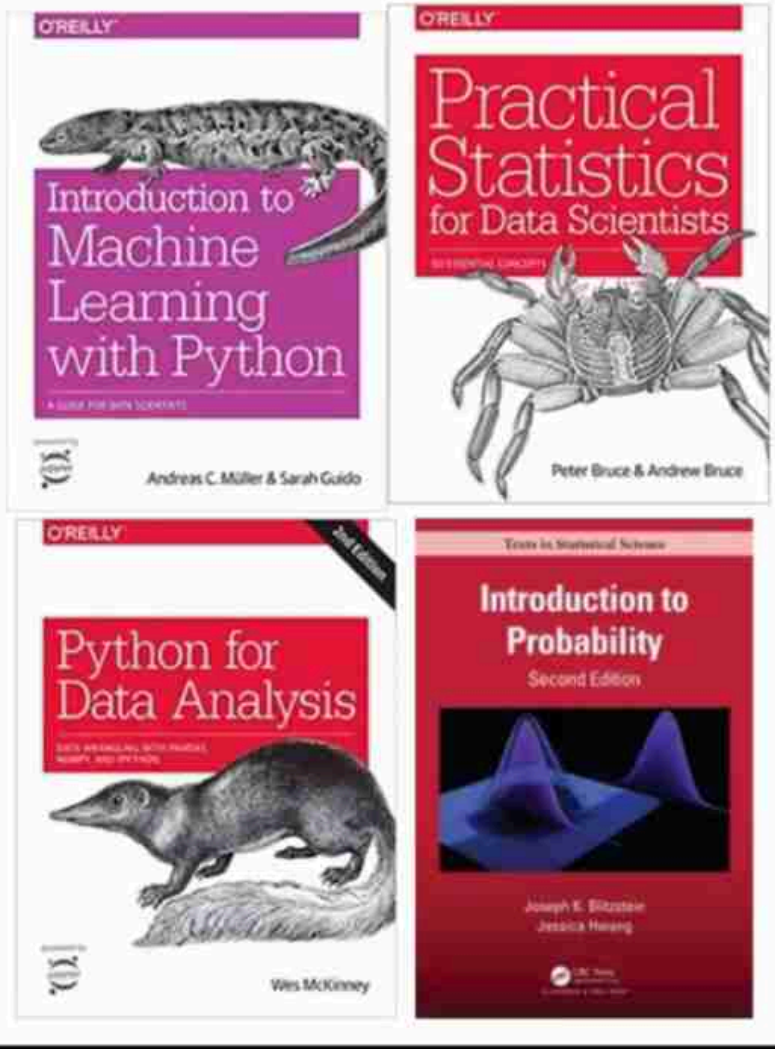
Introduction to Statistics and Data

Ayan Seal

- Course Name: Introduction to Data Science
- Course Code: CS2004
- Books:
- Marks (100)
  - Quiz 1: 10
  - Mid-Semester: 25
  - Quiz 2: 10
  - End-Semester:35
  - Lab: 20



Timetable

- Monday (class): 12 pm to 1 pm
- Tuesday (class): 9 am to 10 am
- Wednesday (lab): 4 pm to 6 pm
- Thursday (class): 10 am to 11 am

What are Statistics?

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory.

Statisticians, people who do statistics, are particularly concerned with determining how to draw reliable conclusions about large groups and general phenomena from the observable characteristics of small samples that represent only a small portion of the large group or a limited number of instances of a general phenomenon.

Cont'd…

- In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Cont'd...

- From the sample data, we can calculate a statistic. A statistic is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A parameter is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

Cont'd…

- **Population**: all math classes
- **Sample**: One of the math classes
- **Parameter**: Average number of points earned per student over all math classes

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics.

Example

- Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

- **Answer:**
  - Population: All the first year college students at ABC College
  - Sample: 100 first year students who are surveyed at the college.
  - Parameter: Average amount of money a first year college student spent on school supplies that do not include books.
  - Statistics: Average amount of money these 100 first year college student spent on school supplies that do not include books.
  - Variable: The amount of money a first year ABC College student spend on school supplies that do not include books.
  - Data: The amount that we collected from these 100 students, like $150, $200, and $225.

Two major areas of statistics

- The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

Cont'd...

Types of Statistics

Descriptive Statistics

Collecting, summarizing, and describing data

Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based only on sample data

Common statistical tools and procedures

- Some common statistical tools and procedures include the following:

  Descriptive

  - Mean (average)
  - Variance
  - Skewness
  - Kurtosis

  Inferential

  - Liner regression analysis
  - Analysis of variance (ANOVA)
  - Logit/Probit models
  - Null hypothesis testing

Example

- If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is 22/40 and the proportion of women students is 18/40. Mean and proportion are discussed in more detail in later chapters.

Descriptive Statistics

Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median, and mode.

Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation.

Cont'd...

The distribution refers to the overall "shape" of the data, which can be depicted on a chart such as a histogram or dot plot, and includes properties such as the probability distribution function, skewness, and kurtosis.

Descriptive statistics can also describe differences between observed characteristics of the elements of a data set. Descriptive statistics help us understand the collective properties of the elements of a data sample and form the basis for testing hypotheses and making predictions using inferential statistics.

Inferential Statistics

Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population from the characteristics of a sample and to decide how certain they can be of the reliability of those conclusions.

Based on the sample size and distribution of the sample data statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

Cont'd...

Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits, or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.

Cont'd...

Regression analysis is a common method of statistical inference that attempts to determine the strength and character of the relationship (or correlation) between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

The output of a regression model can be analyzed for statistical significance, which refers to the claim that a result from findings generated by testing or experimentation is not likely to have occurred randomly or by chance but are instead likely to be attributable to a specific cause elucidated by the data. Having statistical significance is important for academic disciplines or practitioners that rely heavily on analyzing data and research.

What is the difference between descriptive and inferential statistics?

Descriptive statistics are used to describe or summarize the characteristics of a sample or data set, such as a variable's mean, standard deviation, or frequency.

Inferential statistics, in contrast, employs any number of techniques to relate variables in a data set to one another, for example using correlation or regression analysis. These can then be used to estimate forecasts or infer causality.

Who uses statistics?

Statistics are used widely across an array of applications and professions. Any time data are collected and analyzed, statistics are being done. This can range from government agencies to academic research to analyzing investments.

Glossary

- **Average:** also called mean; a number that describes the central tendency of the data
- **Categorical Variable:** variables that take on values that are names or labels
- **Data:** a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)
- **Numerical Variable:** variables that take on values that are indicated by numbers
- **Parameter:** a number that is used to represent a population characteristic and that generally cannot be determined easily

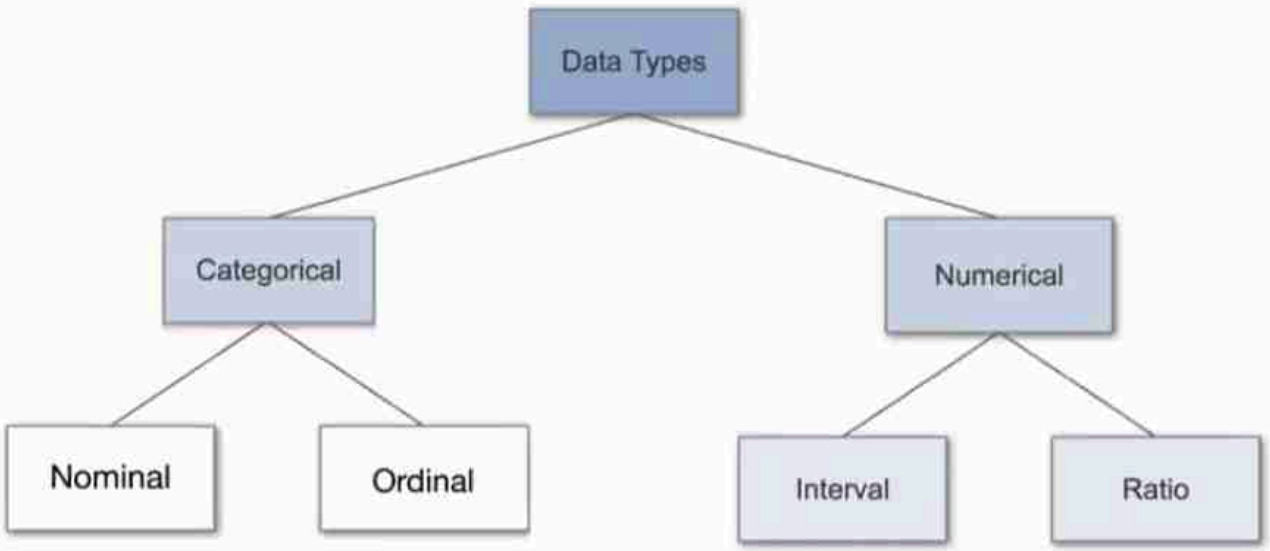
Cont'd...

- **Population:** all individuals, objects, or measurements whose properties are being studied
- **Probability:** a number between zero and one, inclusive, that gives the likelihood that a specific event will occur
- **Proportion:** the number of successes divided by the total number in the sample
- **Representative Sample:** a subset of the population that has the same characteristics as the population
- **Sample:** a subset of the population studied
- **Statistic:** a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.
- **Variable:** a characteristic of interest for each person or object in a population

Data

- **Data** are units of information, often numeric, that are collected through observation. In more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects,[1] while a **datum** (singular of *data*) is a single value of a single variable
- **Data Types are an important concept of statistics, which needs to be understood, to correctly apply statistical measurements to your data and therefore to correctly conclude certain assumptions about it.**
- Having a good understanding of the different data types, also called measurement scales, is a crucial prerequisite for doing Exploratory Data Analysis (EDA), since you can use certain statistical measurements only for specific data types.
- Data science is all about experimenting with raw or structured data. Data is the fuel that can drive a business to the right path or at least provide actionable insights that can help strategize current campaigns, easily organize the launch of new products, or try out different experiments.

Data Types in Statistics



Categorical Data

- Qualitative or Categorical data describes the object under consideration using a finite set of discrete classes. It means that this type of data can't be counted or measured easily using numbers and therefore divided into categories. The gender of a person (male, female, or others) is a good example of this data type.
- These are usually extracted from audio, images, or text medium. Another example can be of a smartphone brand that provides information about the current rating, the color of the phone, category of the phone, and so on. All this information can be categorized as Qualitative data.

Nominal Data

- Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as "labels". Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features below:

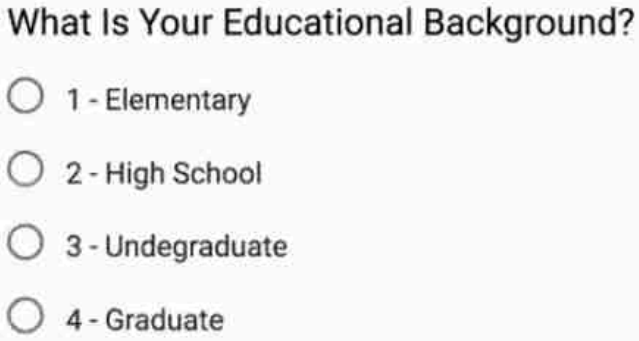
- The left feature that describes if a person is married would be called "dichotomous", which is a type of nominal scales that contains only two categories.

Cont'd...

- These are the set of values that don't possess a natural ordering. Let's understand this with some examples. The color of a smartphone can be considered as a nominal data type as we can't compare one color with others.
- It is not possible to state that 'Red' is greater than 'Blue'. The gender of a person is another one where we can't differentiate between male, female, or others. Mobile phone categories whether it is midrange, budget segment, or premium smartphone is also nominal data type.

Ordinal Data

- Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:



- Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

NOTICES

Notice No.	20210722-3	Notice Date	22 Jul 2021
Category	Company related	Segment	Equity
Subject	Listing of Equity Shares of Zomato Limited		

Content

Trading members of the Exchange are hereby informed that the equity shares of Zomato Limited shall be listed and admitted to dealings on the Exchange in due course.

Name of the company	Scrip Code	Symbol	ISIN No.
Zomato Limited	543320	ZOMATO	INE758T01015

The date of listing and the details of the securities shall be informed through a separate notice.

For and on behalf of BSE Limited

Rupal Khandelwal

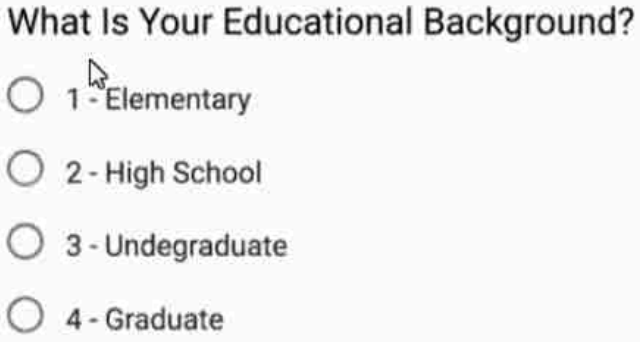
Assistant General Manager

Date: Thursday, July 22, 2021

It's almost final that Zomato is listing tomorrow.

Ordinal Data

- Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:



- Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

Cont'd...

- These types of values have a natural ordering while maintaining their class of values. If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large. The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.
- These categories help us deciding which encoding strategy can be applied to which type of data. Data encoding for Qualitative data is important because machine learning models can't handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.

Numerical Data

- This data type tries to quantify things and it does by considering numerical values that make it countable in nature. The price of a smartphone, discount offered, number of ratings on a product, the frequency of processor of a smartphone, or ram of that particular phone, all these things fall under the category of Quantitative data types.
- The key thing is that there can be an infinite number of values a feature can take. For instance, the price of a smartphone can vary from x amount to any value and it can be further broken down based on fractional values.

Discrete vs. Continuous

- **Discrete**
  - The numerical values which fall under are integers or whole numbers are placed under this category. The number of speakers in the phone, cameras, cores in the processor, the number of sims supported all these are some of the examples of the discrete data type.
- **Continuous**
  - The fractional numbers are considered as continuous values. These can take the form of the operating frequency of the processors, the android version of the phone, wifi frequency, temperature of the cores, and so on.

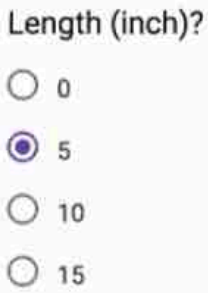
Interval Data

- Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see below:

- The problem with interval values data is that they **don't have a „true zero"**. That means in regards to our example, that there is no such thing as no temperature. With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios. Because there is no true zero, a lot of descriptive and inferential statistics can't be applied.

Ratio Data

- Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero**. Good examples are height, weight, length etc.



Data Visualization

Timetable

- Monday: 2 pm to 3 pm
- Tuesday: 4 pm to 5 pm
- Wednesday (lab): 4 pm to 6 pm
- Thursday: 4 pm to 6 pm

What is Data Visualization?

- Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

What is Visualization and Data Mining?

- **Visualize**: "To form a mental vision, image, or picture of (something not visible or present to the sight, or of an abstraction); to make visible to the mind or imagination."
- **Visualization** is the use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data.
- **Visual Data Mining** is the process of discovering implicit but useful knowledge from large data sets using visualization techniques.

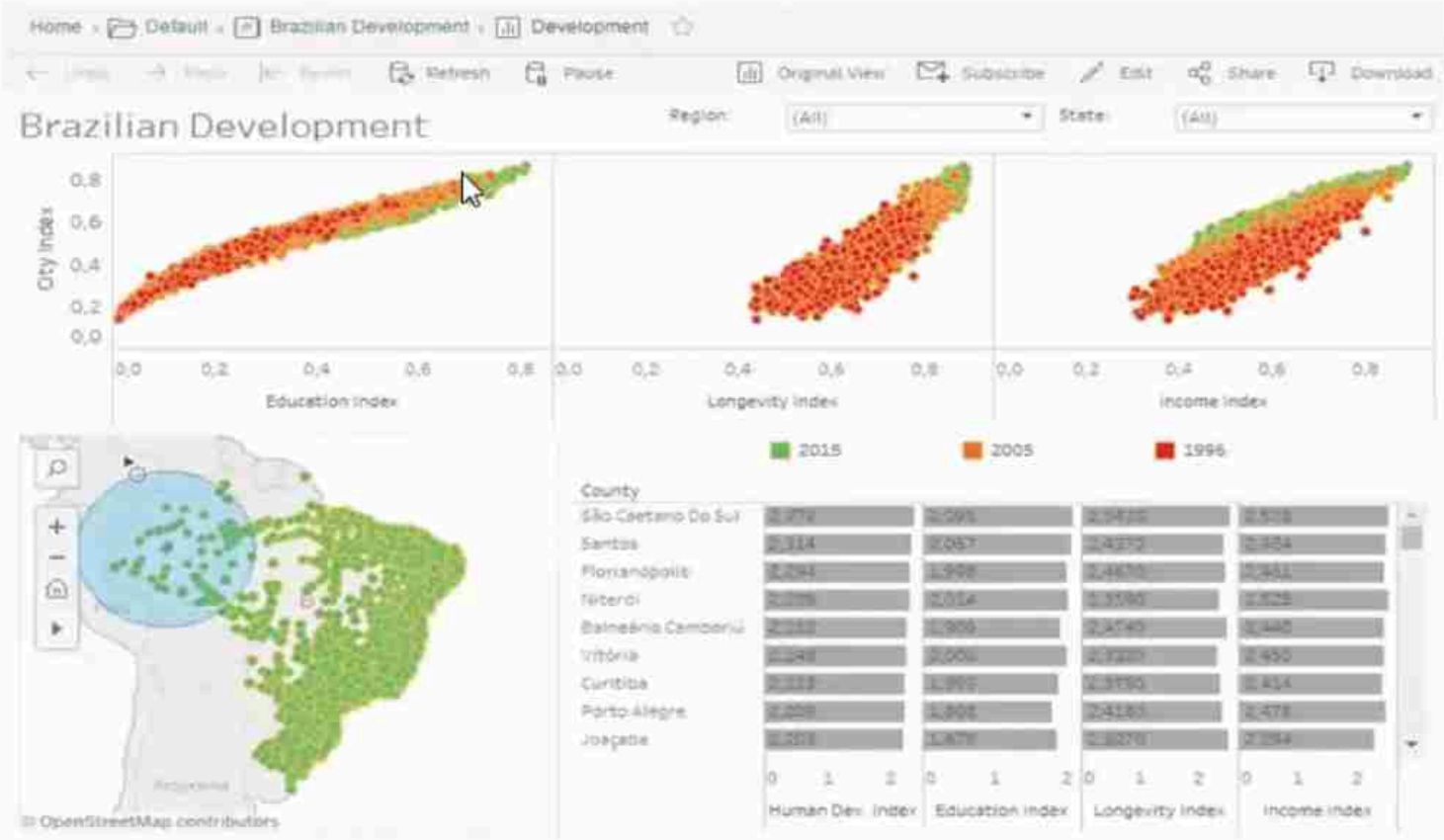
The benefits of Data Visualization

- Our eyes are drawn to colors and patterns. We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies.
- Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

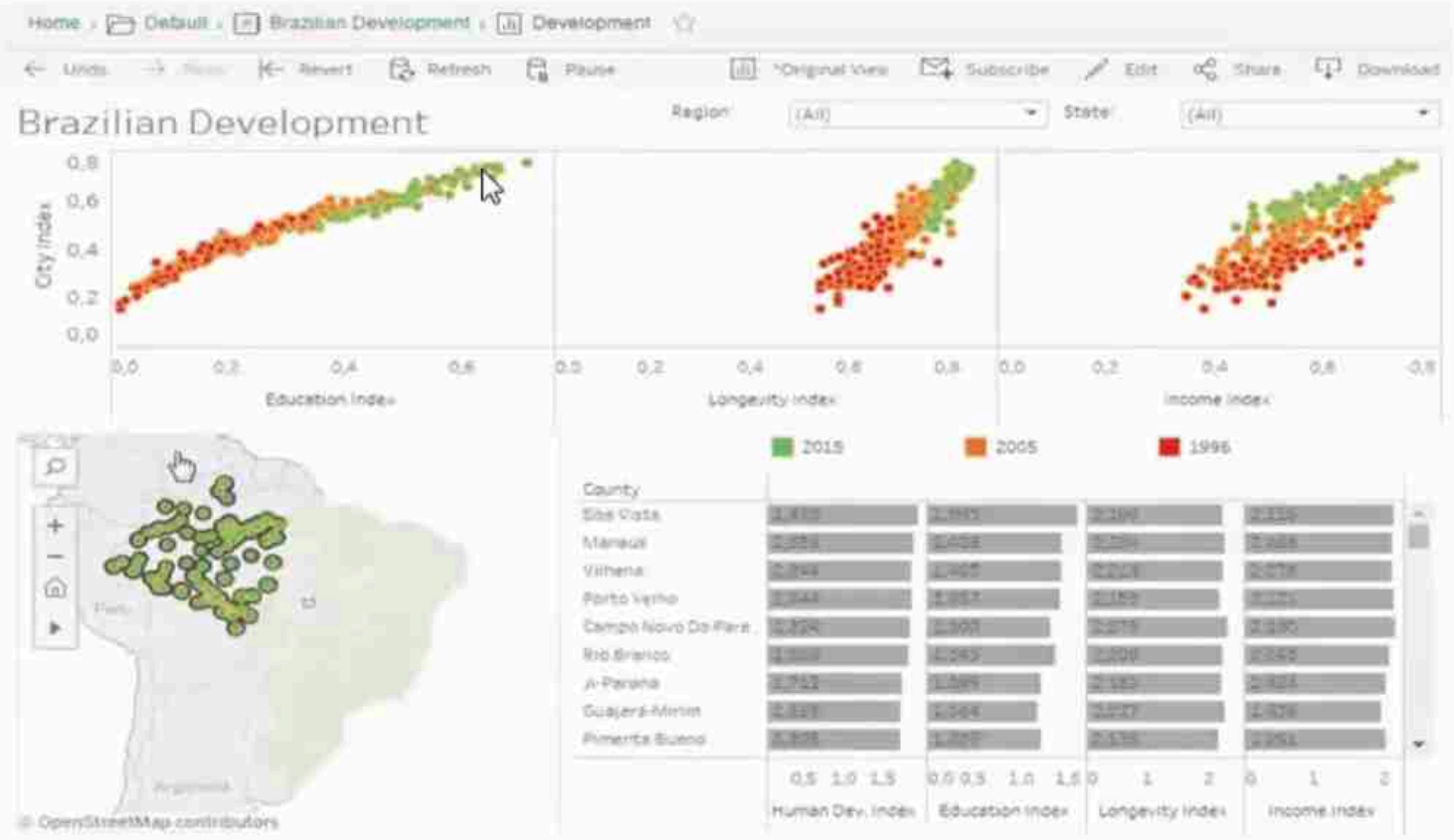
Cont’d…


Brazilian Development
Region
State
City Index
Education Index
Longevity Index
Income Index
2013
2005
1996
County

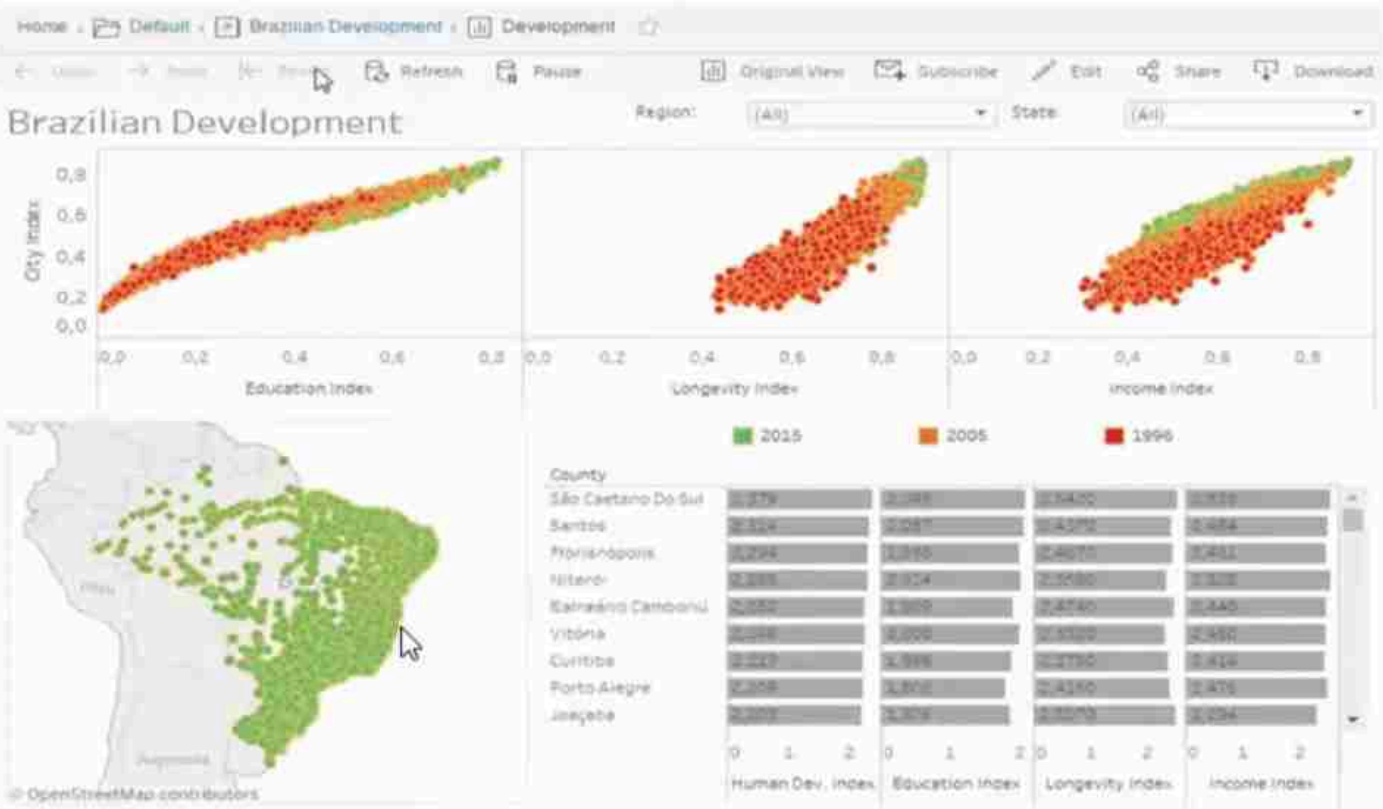
Cont’d…


Home
Default
Brazilian Development
Development
Refresh
Pause
Original View
Subscribe
Edit
Share
Download
Brazilian Development
Region:
(All)
State:
(All)
City Index
0,0
0,2
0,4
0,6
0,8
Education Index
Longevity Index
Income Index
2015
2005
1996
County
São Caetano Do Sul
Santos
Florianópolis
Niterói
Balneário Camboriú
Vitória
Curitiba
Porto Alegre
Joaçaba
Human Dev. Index
Education Index
Longevity Index
Income Index
© OpenStreetMap contributors

Cont’d…


Home
Default
Brazilian Development
Development
Undo
Revert
Refresh
Pause
Original View
Subscribe
Edit
Share
Download
Brazilian Development
Region
(All)
State
(All)
City Index
Education Index
Longevity Index
Income Index
2015
2005
1996
County
Human Dev. Index
Education Index
Longevity Index
Income Index
© OpenStreetMap contributors

Cont’d…

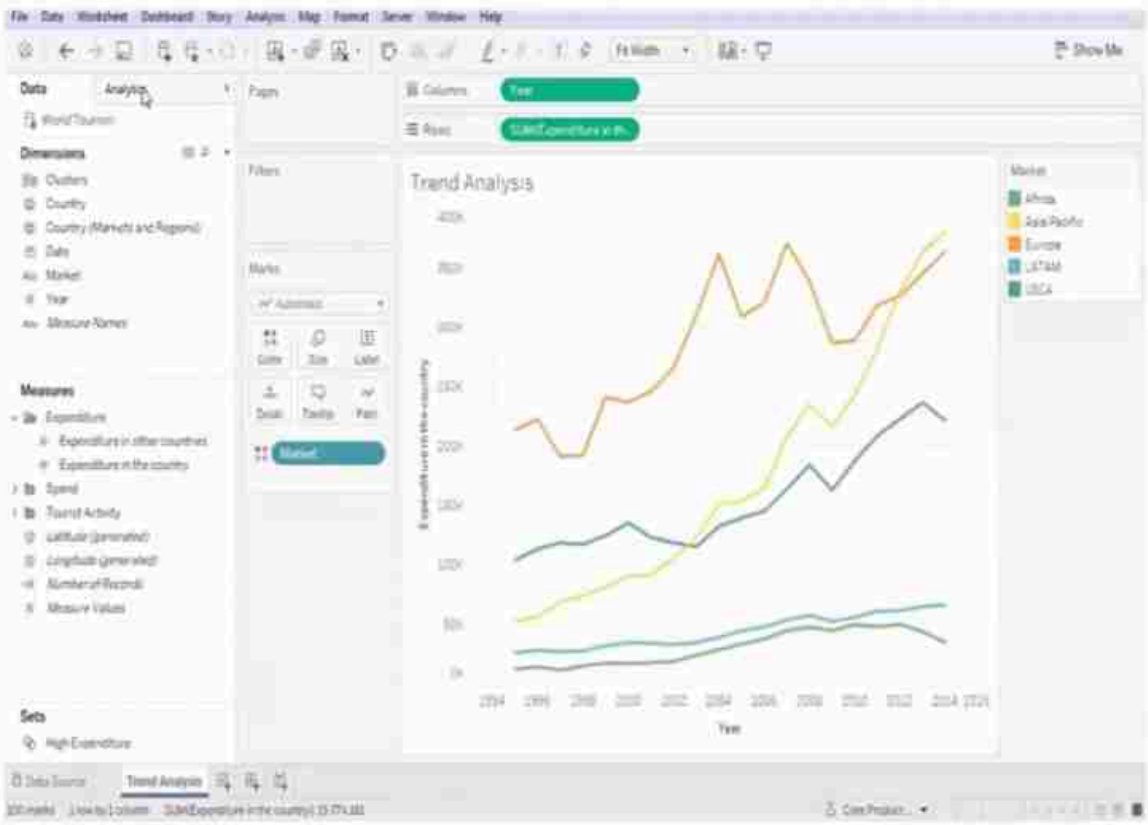

Home
Default
Brazilian Development
Development
Refresh
Pause
Original View
Subscribe
Edit
Share
Download
Brazilian Development
Region:
(All)
State
(All)
City Index
Education Index
Longevity Index
Income Index
2015
2005
1996
County
São Caetano Do Sul
Santos
Florianópolis
Niterói
Balneário Camboriú
Vitória
Curitiba
Porto Alegre
Joaçaba
Human Dev. Index
Education Index
Longevity Index
Income Index
© OpenStreetMap contributors

The different types of visualizations

- When you think of data visualization, your first thought probably immediately goes to simple bar graphs or pie charts. While these may be an integral part of visualizing data and a common baseline for many data graphics, the right visualization must be paired with the right set of information. Simple graphs are only the tip of the iceberg. There's a whole selection of visualization methods to present data in effective and interesting ways.

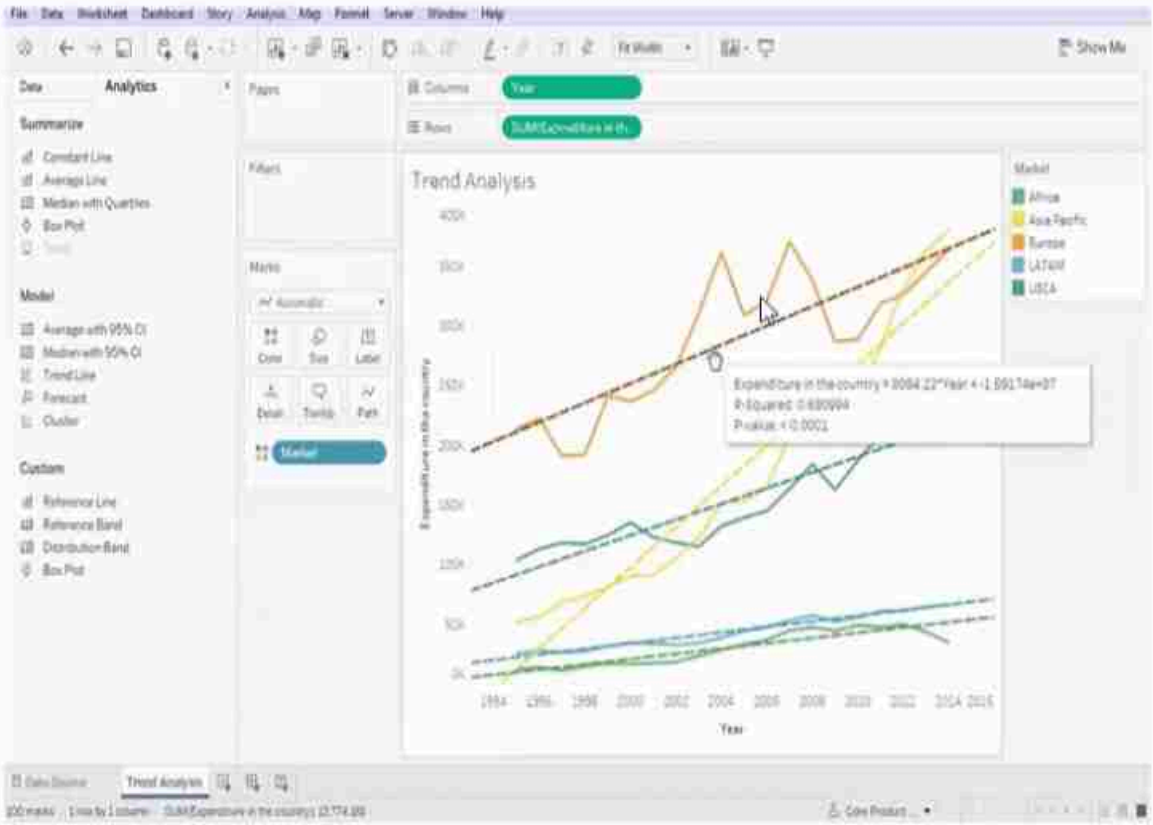
Common types of data visualization

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards



Common types of data visualization

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards



More specific examples of methods to visualize data

- Area Chart
- Bar Chart
- Box-and-whisker Plots
- Bubble Cloud
- Bullet Graph
- Cartogram
- Circle View
- Dot Distribution Map
- Gantt Chart
- Heat Map
- Highlight Table
- Histogram
- Matrix
- Network
- Polar Area
- Radial Tree
- Scatter Plot (2D or 3D)
- Streamgraph
- Text Tables
- Timeline
- Treemap
- Wedge Stack Graph
- Word Cloud
- And any mix-and-match combination in a dashboard!

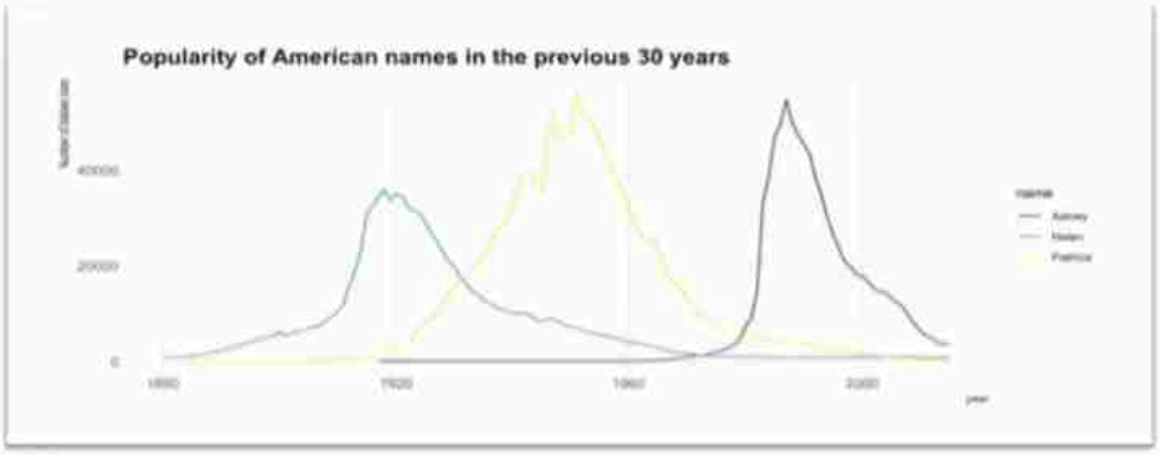
Line Chart

- A line chart or line graph displays the evolution of one or several numeric variables. Data points are connected by straight line segments. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically.

The following example shows the evolution of the bitcoin price between April 2013 and April 2018.

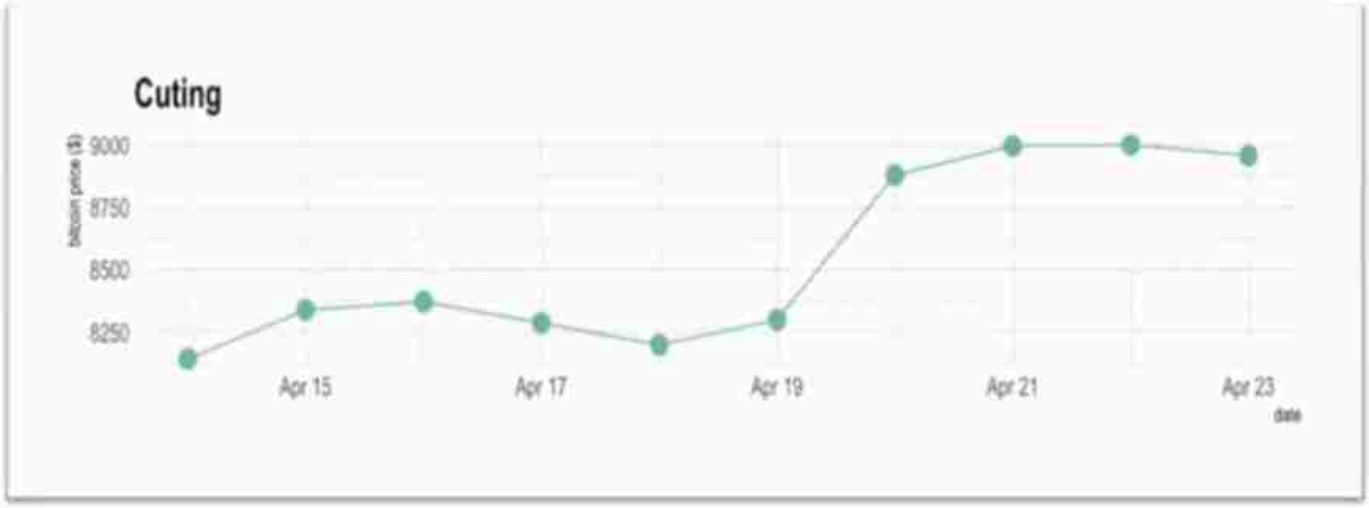
Cont'd...

- Line chart can be used to show the evolution of one (like above) or several variables. Here is an example showing the evolution of three baby name frequencies in the US between 1880 and 2015. Note that this works well for a low number of group to display. With more than a few, the graphic get cluttered and becomes unreadable. This is called a spaghetti chart.



Variation

- If the number of data points is low, it is advised to represent each individual observation with a dot. It allows to understand when exactly the observation have been made:

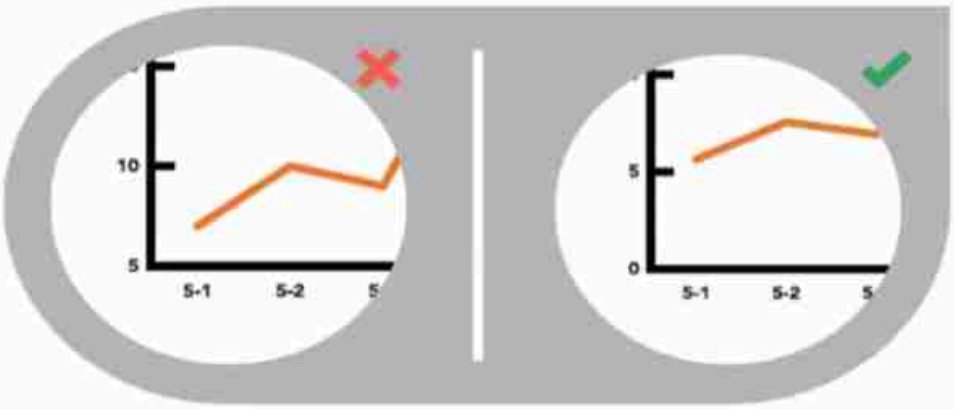


When to Use a Line Chart

- If seeing the trend of your data is the goal, then this is the chart to use. Line charts show time-series relationships using continuous data (that which is measured and has a value within a range). They allow a quick assessment of acceleration (lines curving upward), deceleration (lines curving downward), and volatility (up/down frequency). They are excellent for tracking multiple data sets on the same chart to see any correlation in trends.

Common caveats

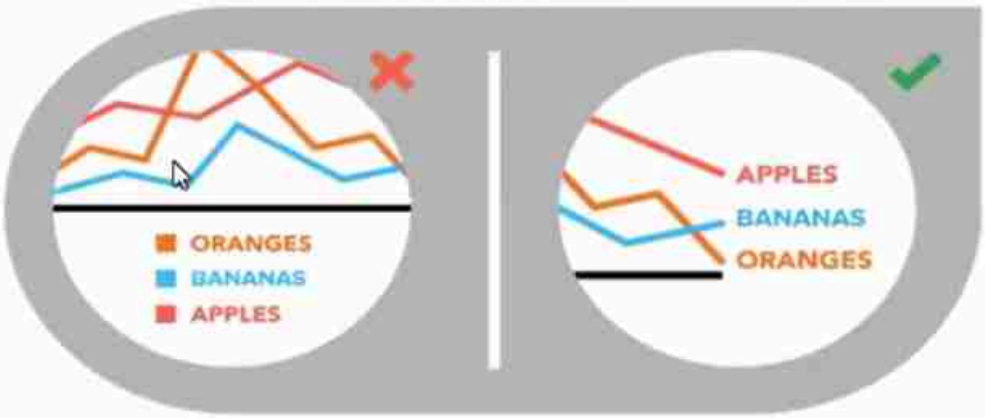
- **Include a Zero Baseline if Possible**



- Although a line chart does not have to start at a zero baseline, it should be included if possible. If relatively small fluctuations in data are meaningful (e.g., stock market data), you may truncate the scale to showcase these variances.

Cont'd...

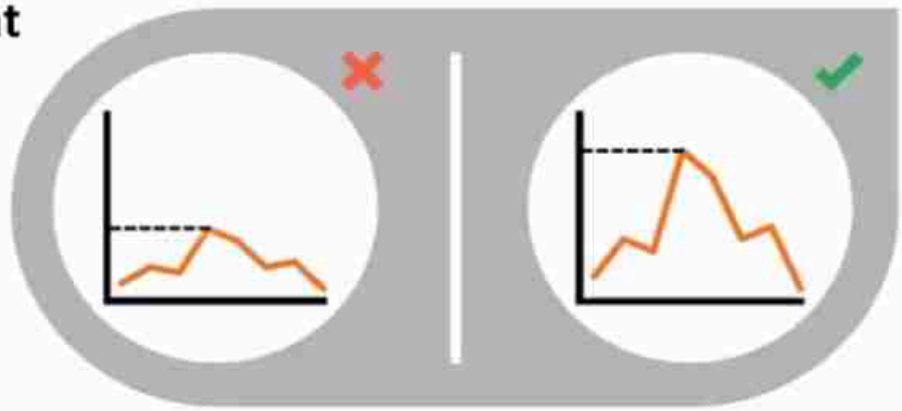
- **Label Lines Directly**



This lets readers quickly identify lines and corresponding labels instead of referencing a legend.

Cont'd...

- **Use the Right Height**

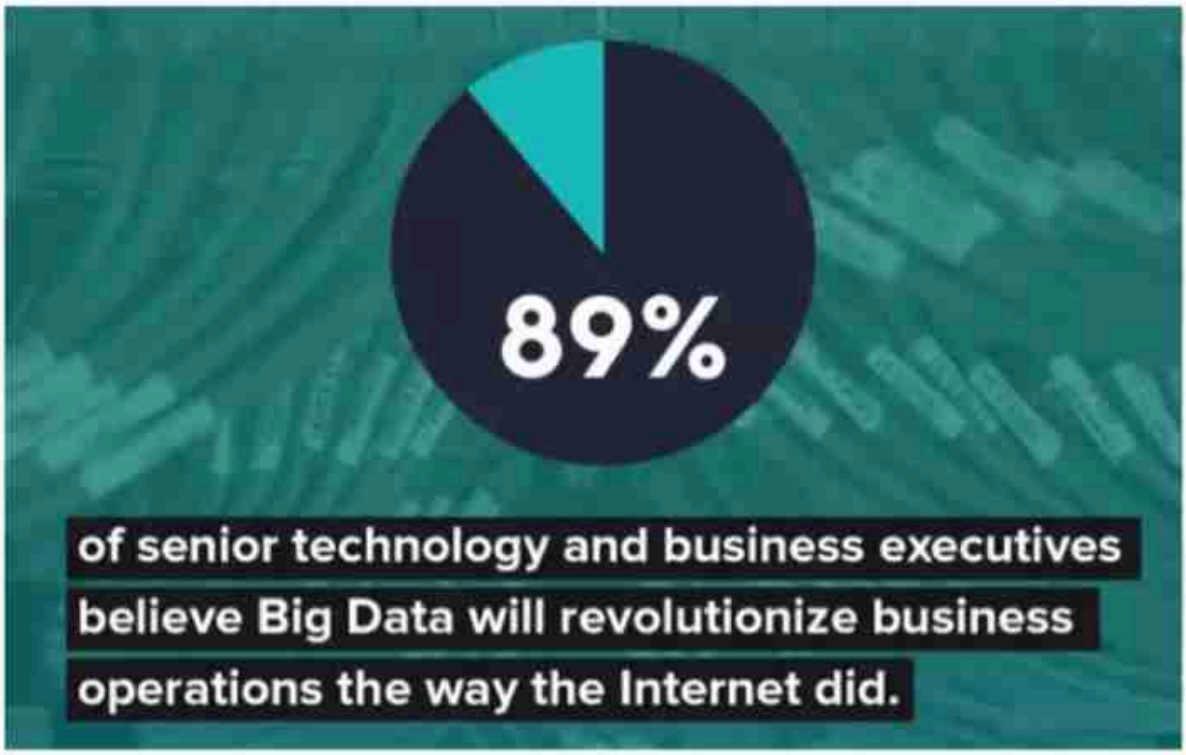


- If you need to compare the evolution of 2 different variables, do not use dual axis. Indeed dual axis can show very different results depending on what range you apply to the axis.
- Mind the spaghetti chart: too many lines make the chart unreadable.
- Think about the aspect ratio of the graphic, extreme ratio make the chart unreadable.

Pie Chart

- Pie charts are one of the oldest and most popular ways to visualize data. This classic chart is the perfect example of the power of data visualization: a simple, easy-to-understand presentation that helps readers instantly identify the parts of a whole.
- The typical pie chart is divided into sections that illustrate a numerical proportion. Each section expresses its quantity through the size of the central angle in proportion to the others. What makes this visually powerful is that the central angle, the outside arc length, and the area these define all proportionally correspond to the quantity they represent.

Cont'd...

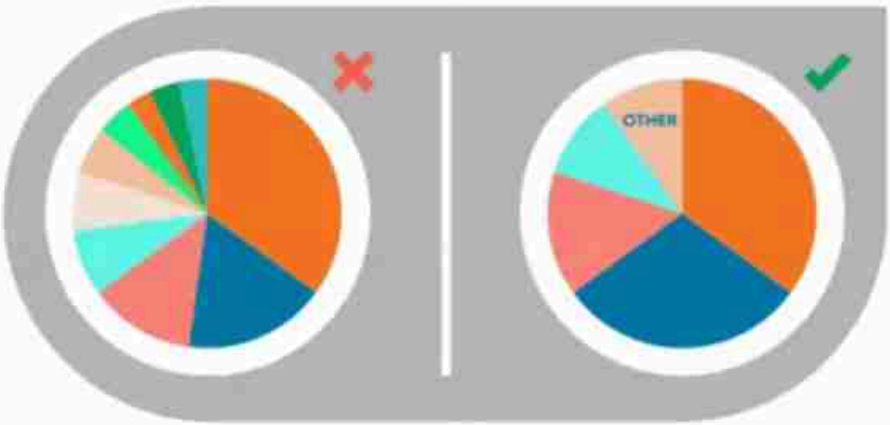

89%
of senior technology and business executives
believe Big Data will revolutionize business
operations the way the Internet did.

When to Use a Pie Chart

- Pie charts are best used when making part-to-whole comparisons (for example, the number of red cars produced each year compared to other popular colors). The message that we are observing fractions of a whole amount is built right into their design—something that can't be said for most other chart types. They have the most impact when the proportion being expressed holds more importance than the specific numbers. They are most clearly understood when using small data sets, often grouping smaller data into an "other" category on the chart.

Tips for Creating Pie Charts

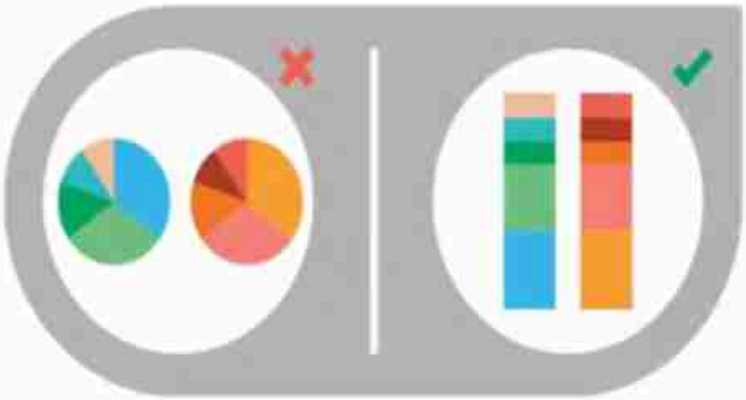
- **Visualize no more than 5 categories per chart**



- It is difficult to differentiate between small values; depicting too many slices decreases the impact of the visualization. If needed, you can group smaller values into an “other” or “miscellaneous” category, but make sure it does not hide interesting or significant information.

Cont'd...

- **Don't use multiple pie charts for comparison**



Slice sizes are very difficult to compare side-by-side. Use a stacked bar chart instead.

Cont'd...

- **Make sure all data adds up to 100%**

Verify that values total 100% and that pie slices are sized proportionately to their corresponding value.

Cont'd...

- **Order slices correctly**

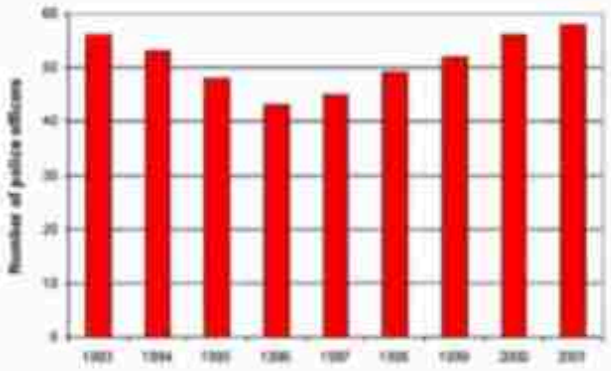
There are two ways to order sections, both of which are meant to aid comprehension:

- **Option 1:** Place the largest section at 12 o'clock, going clockwise. Place the second largest section at 12 o'clock, going counterclockwise. The remaining sections can be placed below, continuing counterclockwise.
- **Option 2:** Start the largest section at 12 o'clock, going clockwise. Place remaining sections in descending order, going clockwise.

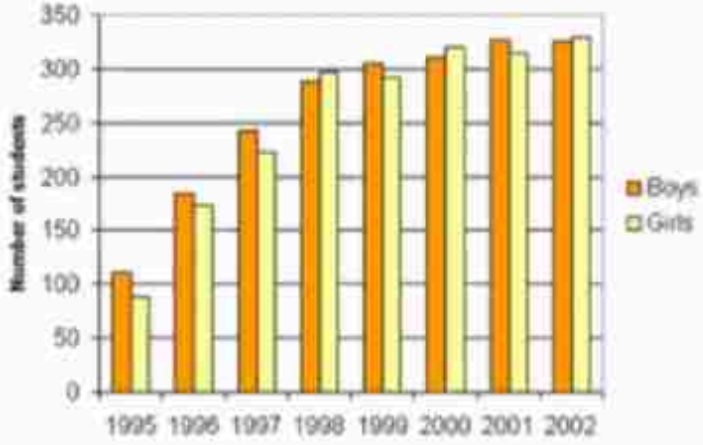
Bar Chart

- Bar charts are a highly versatile way to visually communicate data. Decidedly straightforward, they can convey the message behind the numbers with impact and meaningful clarity, making complex data easy to understand at a glance.
- The bar chart is a chart with rectangular columns proportional in length to the values they represent. Simply put, longer bars equal bigger numbers. On one axis these bars compare categories, while on the other they represent a discrete value.

Example



Number of police officers



Internet use at a school

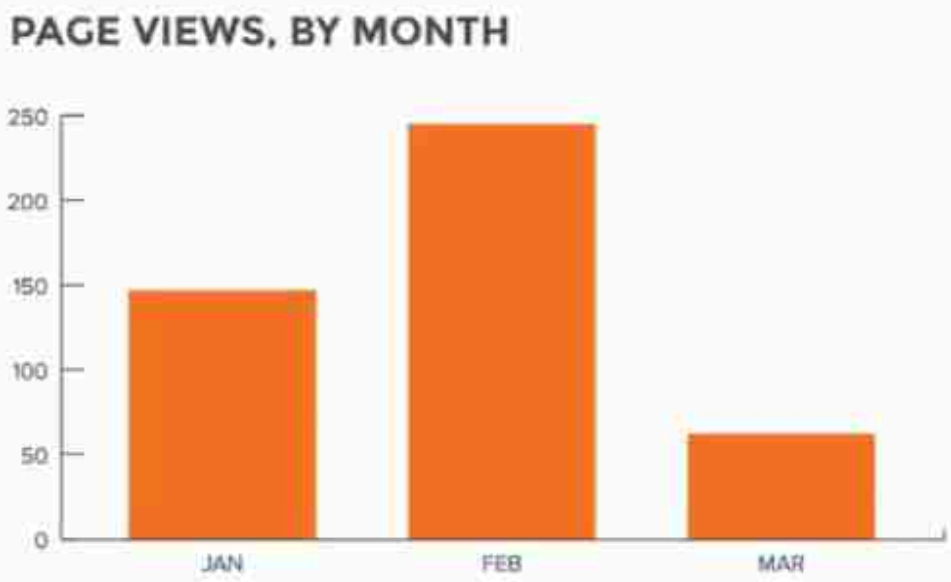
When to Use

Bar charts are used to showcase discrete data—the data that is based on counts and can only be certain values. They are most effectively to:

- Show change over time (e.g., the net monthly earnings of Tesla Motors in a year)
- Compare values of different categories (e.g., a year's fishing yields for different species of fish)
- Compare parts of a whole (e.g., the percent distribution of Netflix rentals across genres)

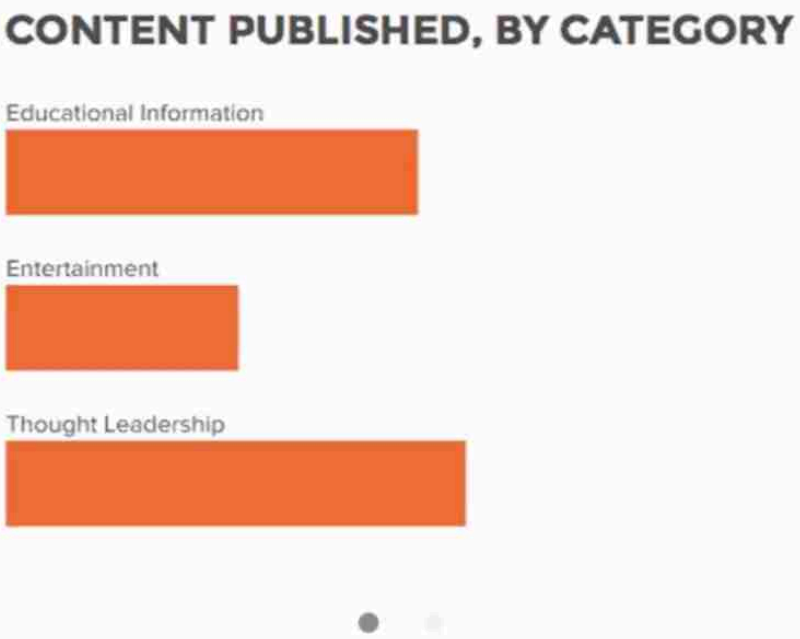
Variations

- The standard bar chart can be used with bars aligned either vertically or horizontally. Vertical orientations are best used for chronological data or when negative values are involved.



Cont'd...

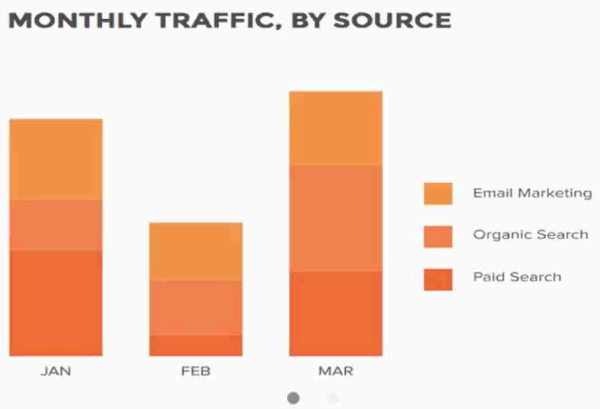
- Horizontal orientations are best used when charting different categories, especially with long labels.



REC

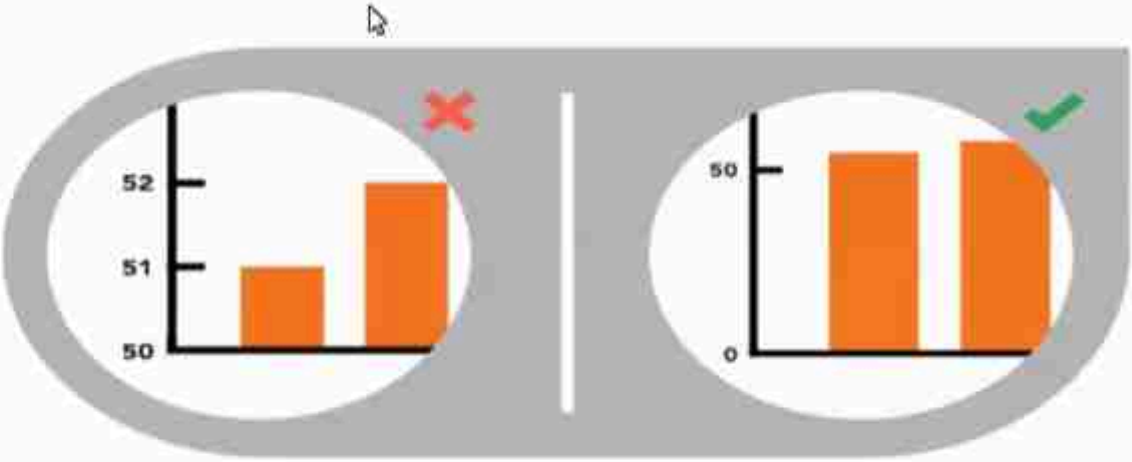
Cont'd...

- Stacked bar charts allow you to showcase and compare part-to-whole relationships.



Best Practices for Designing Bar Charts

- **Start the Y-Axis at 0**



Starting at a value above zero truncates the bars and doesn't accurately reflect the full value

Cont'd...

- **Space bars appropriately**

The space between bars should be roughly ½ bar width.

Cont'd...

- **Use consistent colors**

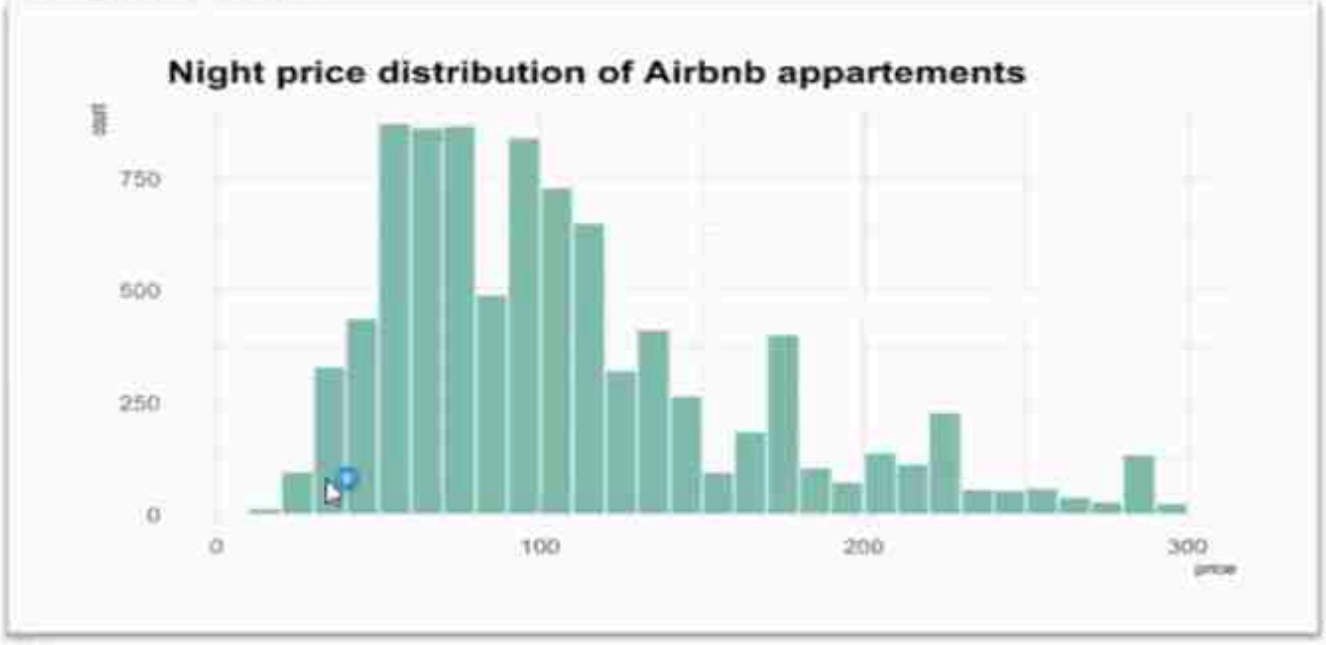
- Use one color for bar charts. You may use an accent color to highlight a significant data point.

Histogram

- A histogram is an accurate graphical representation of the distribution of a numeric variable. It takes as input numeric variables only. The variable is cut into several bins, and the number of observation per bin is represented by the height of the bar.

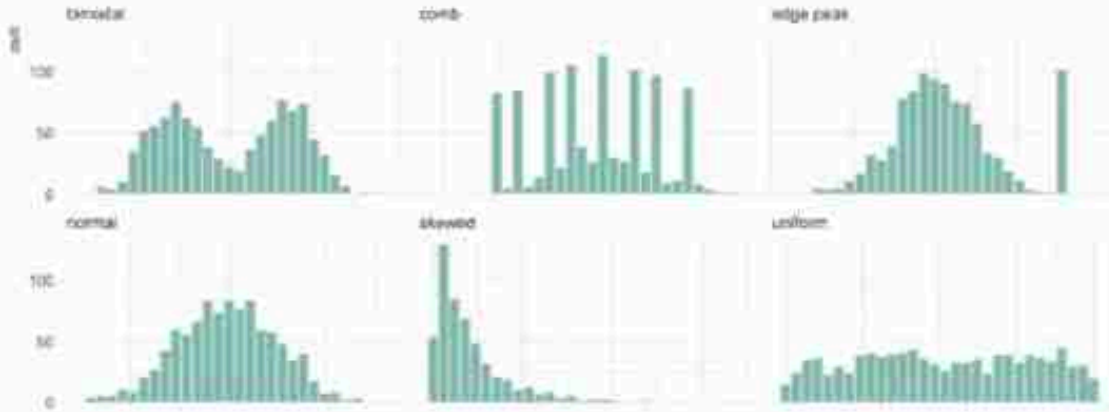
Example

- Here is an example showing the distribution of the night price of Airbnb appartements in the south of France. Price range is divided per 10 euros interval. For example, there are slightly less than 750 appartements with a night price between 100 and 110 euros:



What for

- Histogram are used to study the distribution of one or a few variables. Checking the distribution of your variables one by one is probably the first task you should do when you get a new dataset. It delivers a good quantity of information. Several distribution shapes exist, here is an illustration of the 6 most common ones:



- Checking this distribution also helps you discovering mistakes in the data. For example, the comb distribution can often denote a rounding that has been applied to the variable or another mistake.

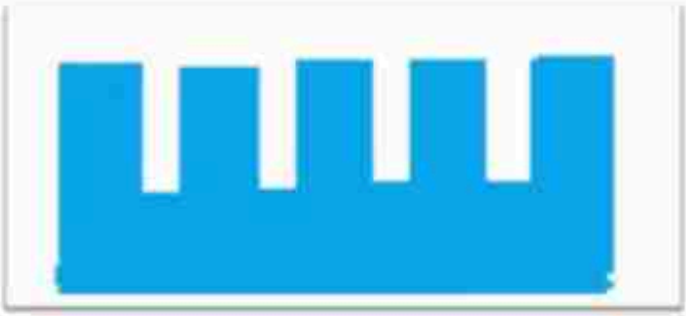
Multimodal Distribution

A multimodal distribution is a probability distribution with more than one peak, or "mode".

- A distribution with one peak is called unimodal.
- A distribution with two peaks is called bimodal.
- A distribution with two peaks or more is multimodal.

Cont'd...

- A **comb distribution** is so-called because the distribution looks like a comb, with alternating high and low peaks. A comb shape can be caused by rounding off. For example, if you are measuring water height to the nearest 10 cm and your class width for the histogram is 5 cm, this could cause a comb shape.

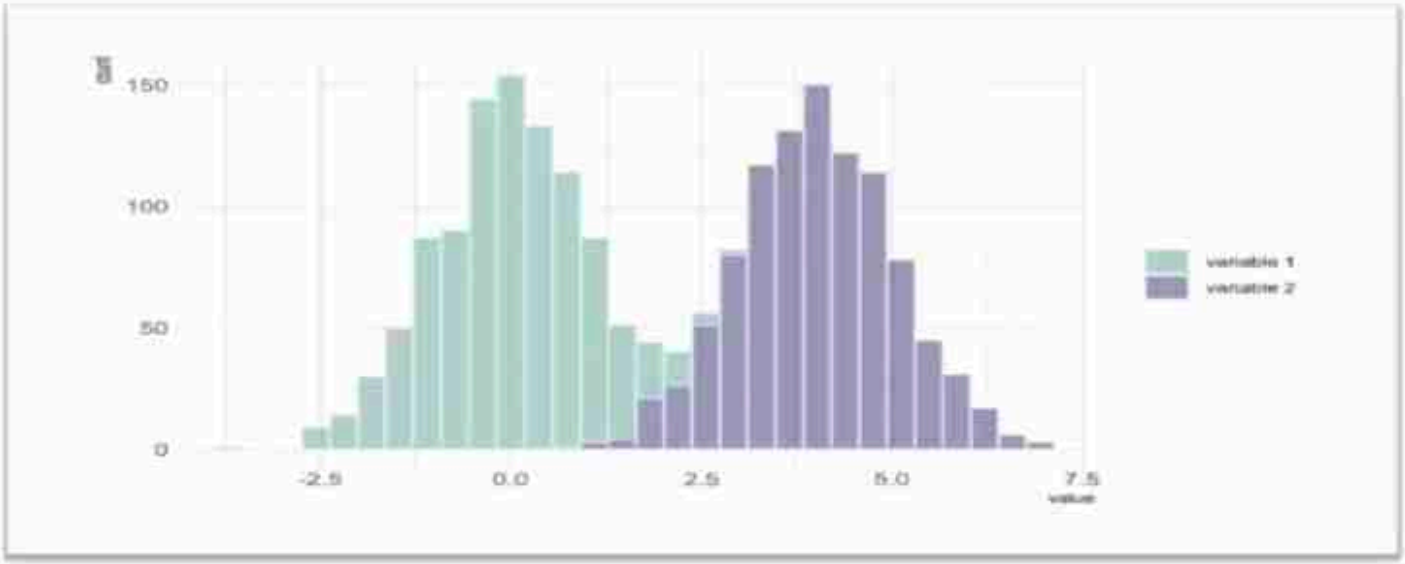


Cont'd...

- A multimodal distribution is known as a **Plateau Distribution** when there are more than a few peaks close together.

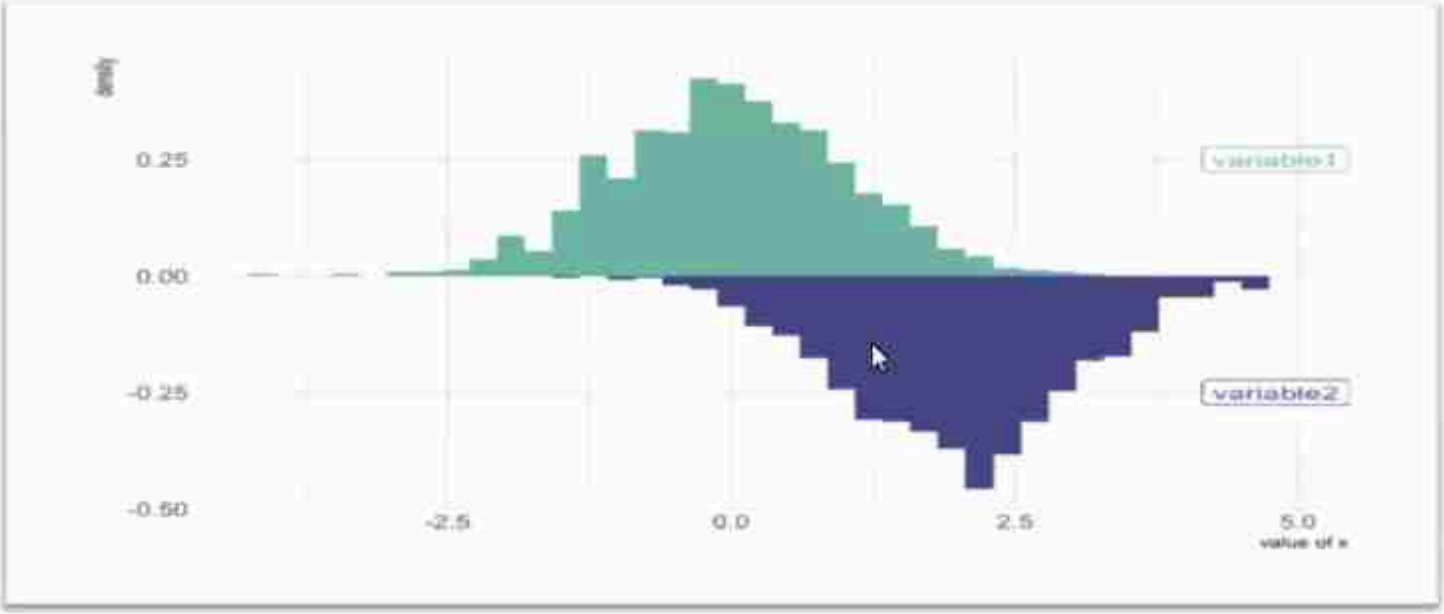
What for

- As a second step, histogram allow to compare the distribution of **a few** variables. Don't compare more than 3 or 4, it would make the figure cluttered and unreadable. This comparison can be done showing the 2 variables on the same graphic and using transparency.



Variation

- A common variation of the histogram is the mirror histogram: it puts face to face 2 histograms to compare their distribution.



Common mistakes

- Try several bin size, it can lead to very different conclusions.
- Don't use weird color scheme. It does not give any more insight.
- Don't confound it with a barplot. A barplot gives a value for each group of a categorical variable. Here, we have only a numeric variable and we check its distribution.
- Don't compare more than 3 groups in the same histogram. The graphic gets cluttered and hardly understandable. Instead use a violin plot, a boxplot, a ridgeline plot or use small multiple.
- Using unequal bin widths.

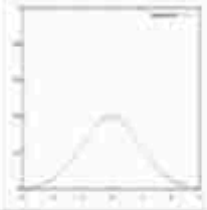
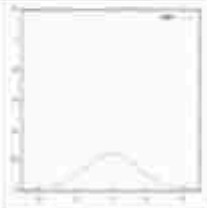
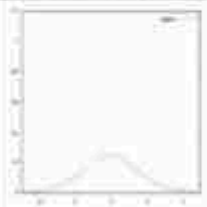
What is a Kernel?

A kernel is a special type of probability density function (PDF) with the added property that it must be even. Thus, a kernel is a function with the following properties

- non-negative
- real-valued
- even
- its definite integral over its support set must equal to 1

Kernel Functions

Uniform ("rectangular window")	$K(u) = \frac{1}{2}$ Support $\lvert u \rvert \leq 1$	 "Boxcar function"
Triangular	$K(u) = (1 - \lvert u \rvert)$ Support $\lvert u \rvert \leq 1$	
Epanechnikov (parabolic)	$K(u) = \frac{3}{4}(1 - u^2)$ Support $\lvert u \rvert \leq 1$	

Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	
Cosine	$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right)$ Support $\lvert u \rvert \leq 1$	
Logistic	$K(u) = \frac{1}{e^u + 2 + e^{-u}}$	
Sigmoid function	$K(u) = \frac{2}{\pi} \frac{1}{e^u + e^{-u}}$	

What is Kernel Density Estimation?

Kernel density estimation is a non-parametric method of estimating the probability density function (PDF) of a continuous random variable. It is non-parametric because it does not assume any underlying distribution for the variable. Essentially, at every datum, a kernel function is created with the datum at its centre – this ensures that the kernel is symmetric about the datum. The PDF is then estimated by adding all of these kernel functions and dividing by the number of data to ensure that it satisfies the 2 properties of a PDF:

- Every possible value of the PDF (i.e. the function,), is non-negative.
- The definite integral of the PDF over its support set equals to 1.

Constructing a Kernel Density Estimate

1) Choose a kernel; the common ones are normal (Gaussian), uniform (rectangular), and triangular.

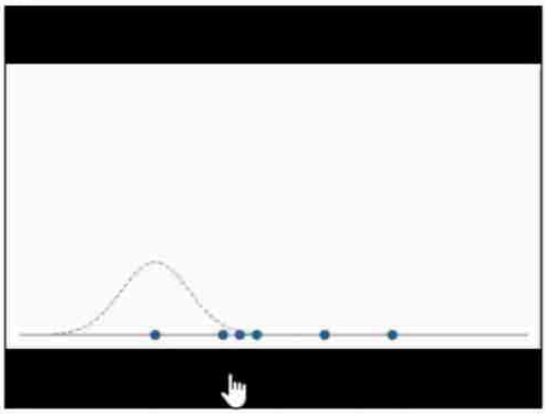
2) At each datum, x_i, build the scaled kernel function

$$h^{-1}K[(x - x_i)/h]$$

where $K()$ is your chosen kernel function. The parameter h is called the bandwidth, the window width, or the smoothing parameter.

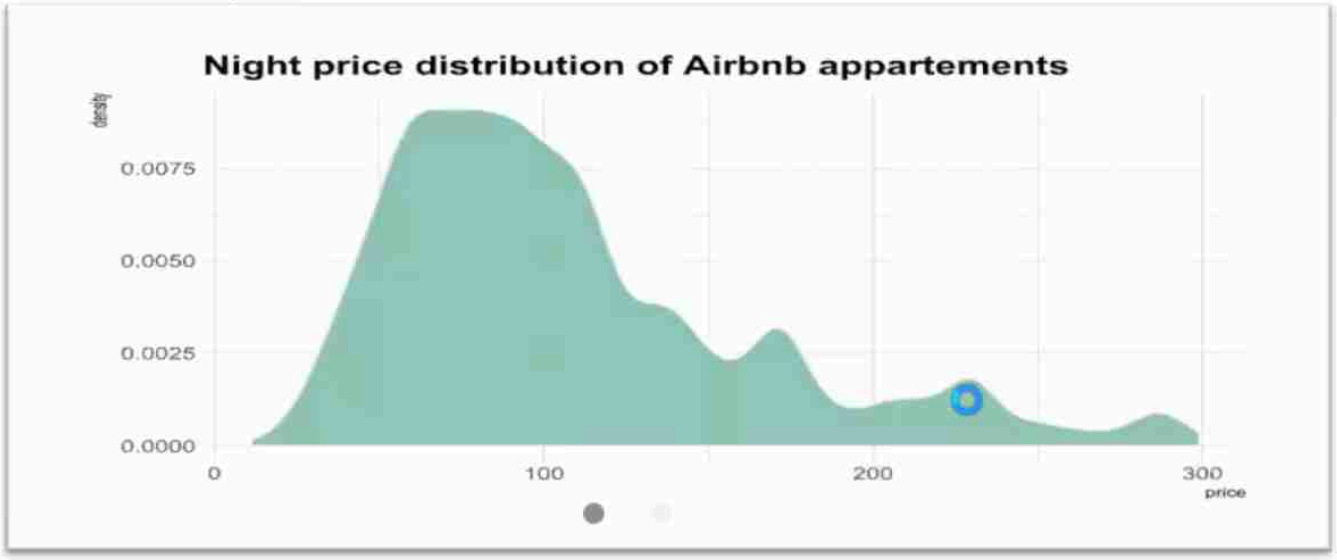
3) Add all of the individual scaled kernel functions and divide by n; this places a probability of $1/n$ to each x_i. It also ensures that the kernel density estimate integrates to 1 over its support set.

$$\hat{f}(x) = n^{-1}h^{-1}\sum_{i=1}^{n} K[(x - x_i)/h]$$



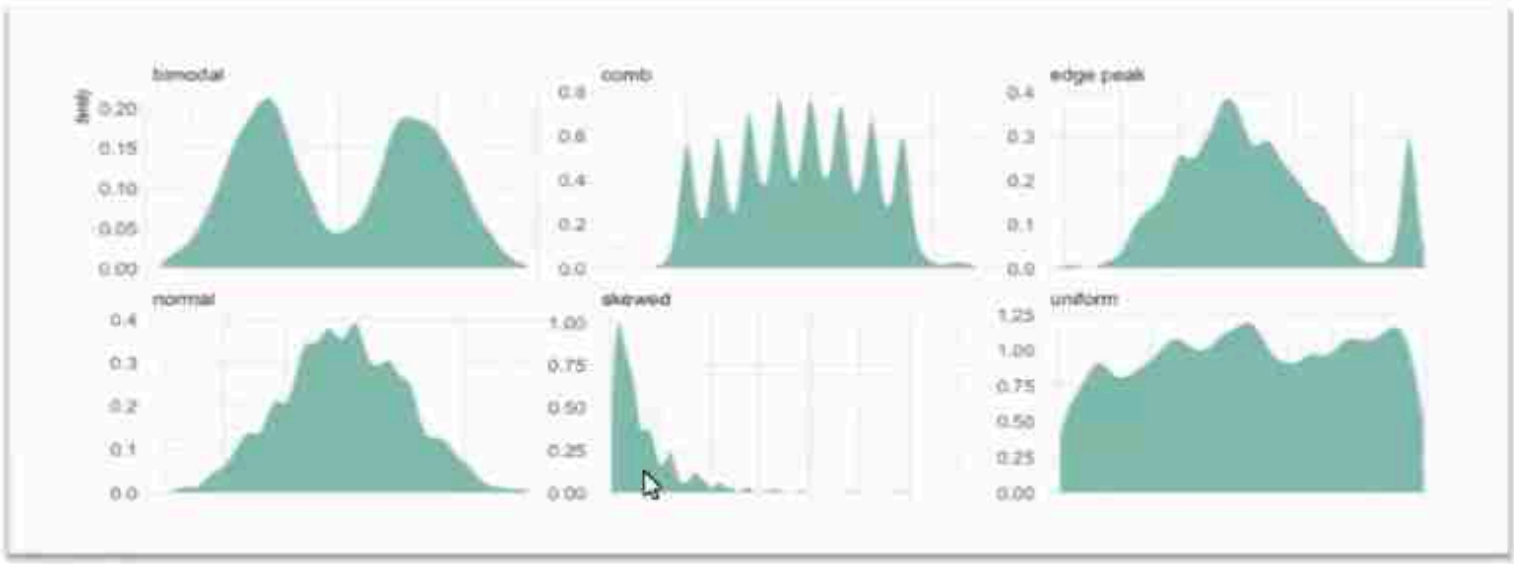
Density

- A density plot is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable.
- It is a smoothed version of the histogram and is used in the same concept.



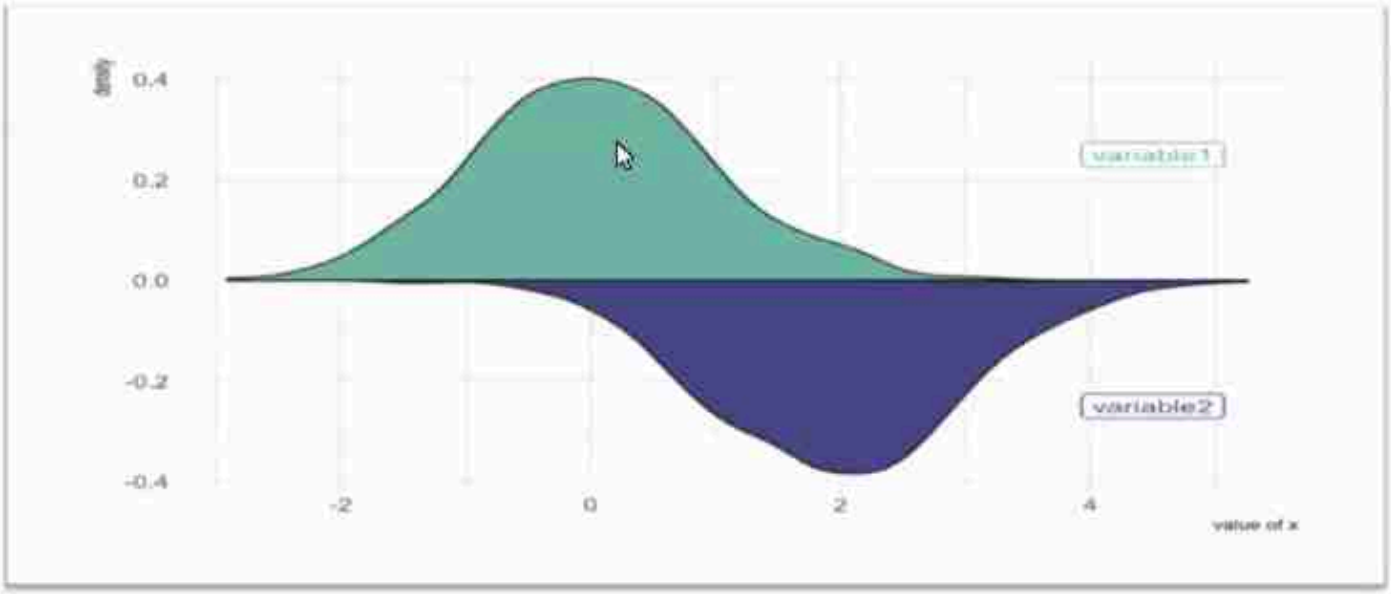
What for

- Density plots are used to study the distribution of one or a few variables. Checking the distribution of your variables one by one is probably the first task you should do when you get a new dataset. It delivers a good quantity of information. Several distribution shapes exist, here is an illustration of the 6 most common ones:



Variation

- A common variation of the histogram is the mirror histogram: it puts face to face 2 histograms to compare their distribution.



Common mistakes

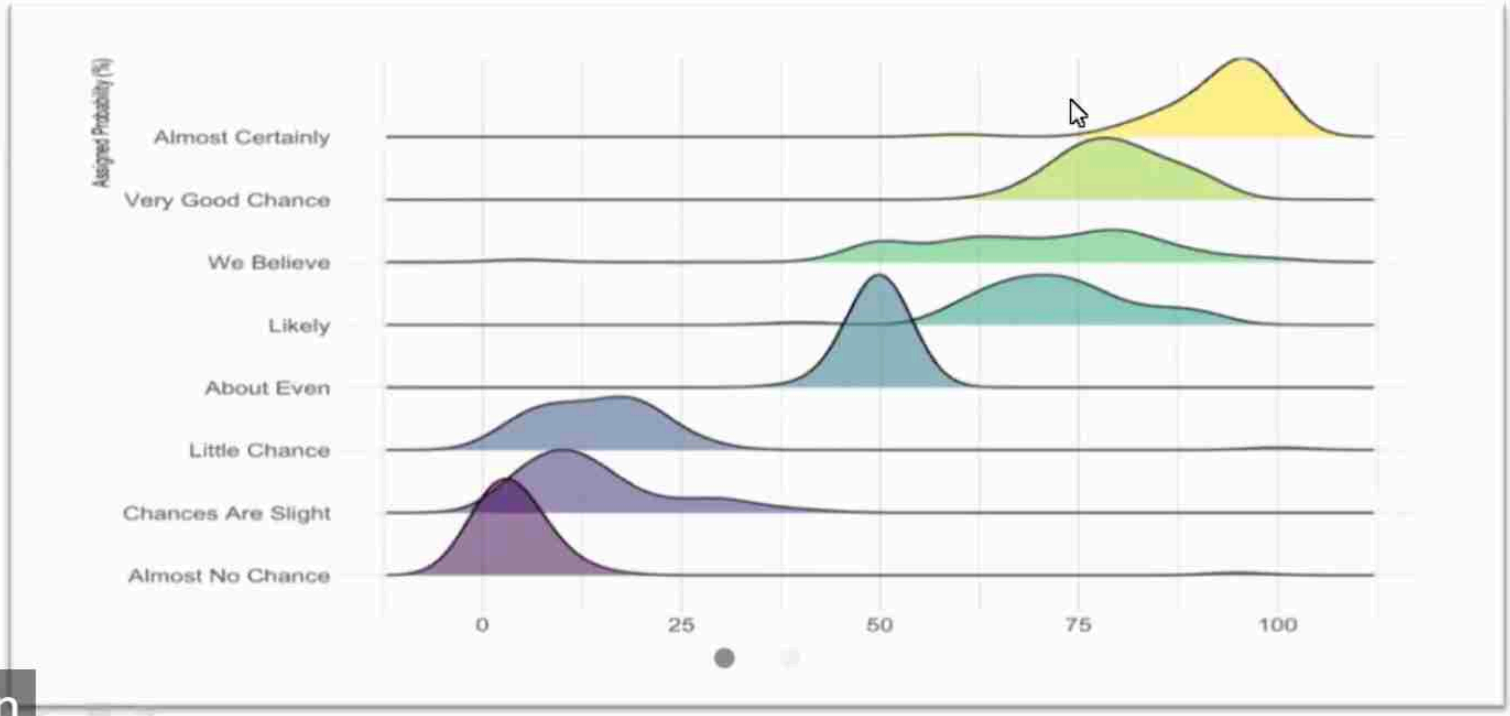
- Play with the bandwidth argument, it can lead to very different conclusions.
- Don't compare more than 3 groups on the same density plot. The graphic gets cluttered and hardly understandable. Instead use a violin plot, a boxplot, a ridgeline plot or use small multiple.

Ridgeline Plot

- A Ridgeline plot (sometimes called Joyplot) shows the distribution of a numeric value for several groups. Distribution can be represented using histograms or density plots, all aligned to the same horizontal scale and presented with a slight overlap.

Example

- *What probability would you assign to the phrase "Highly likely"* were asked. Answers between 0 and 100 were recorded, and here is the distribution for each question:

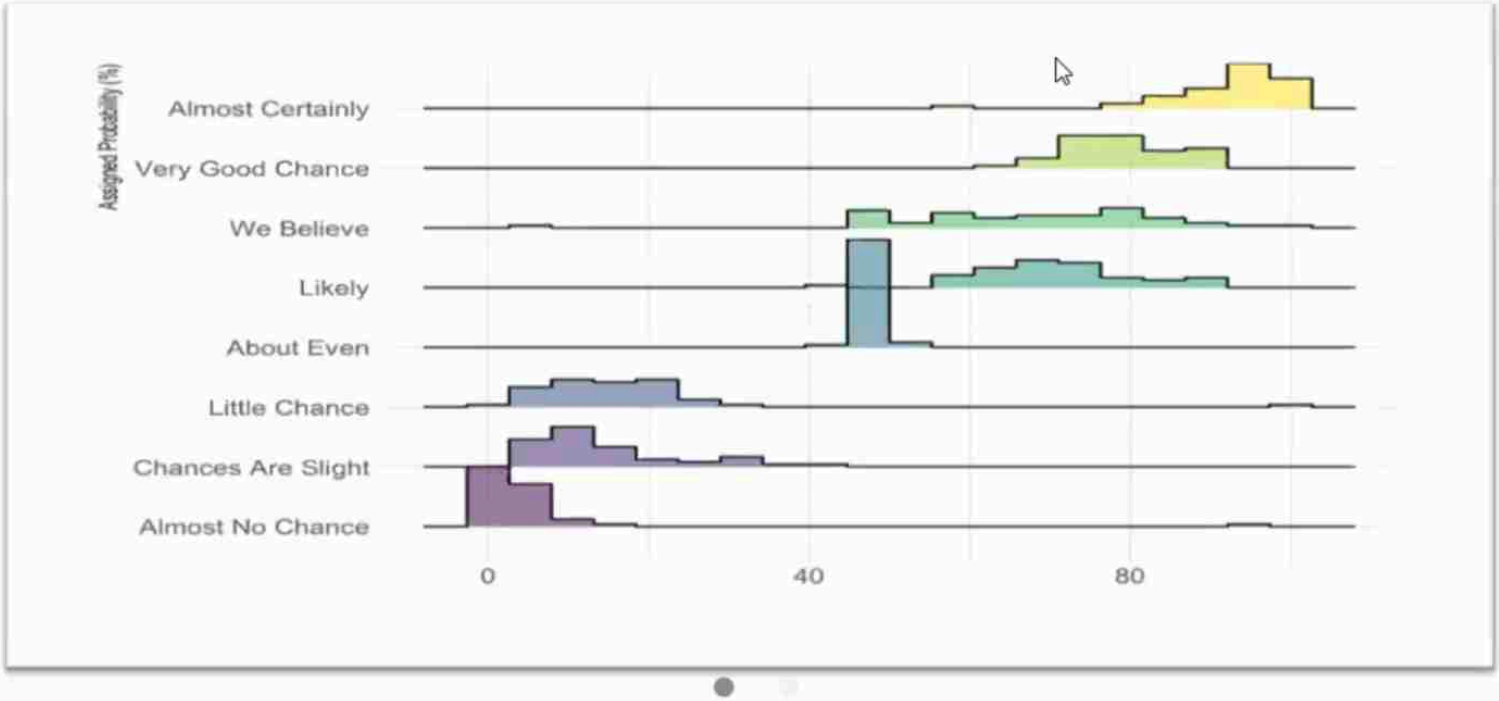


What for

- Ridgeline plots make sense when the number of group to represent is medium to high, and thus a classic window separation would take to much space. Indeed, the fact that groups overlap each other allows to use space more efficiently. If you have less than 6 groups, dealing with other distribution plots is probably better.
- It works well when there is a clear pattern in the result, like if there is an obvious ranking in groups. Otherwise group will tend to overlap each other, leading to a messy plot not providing any insight.

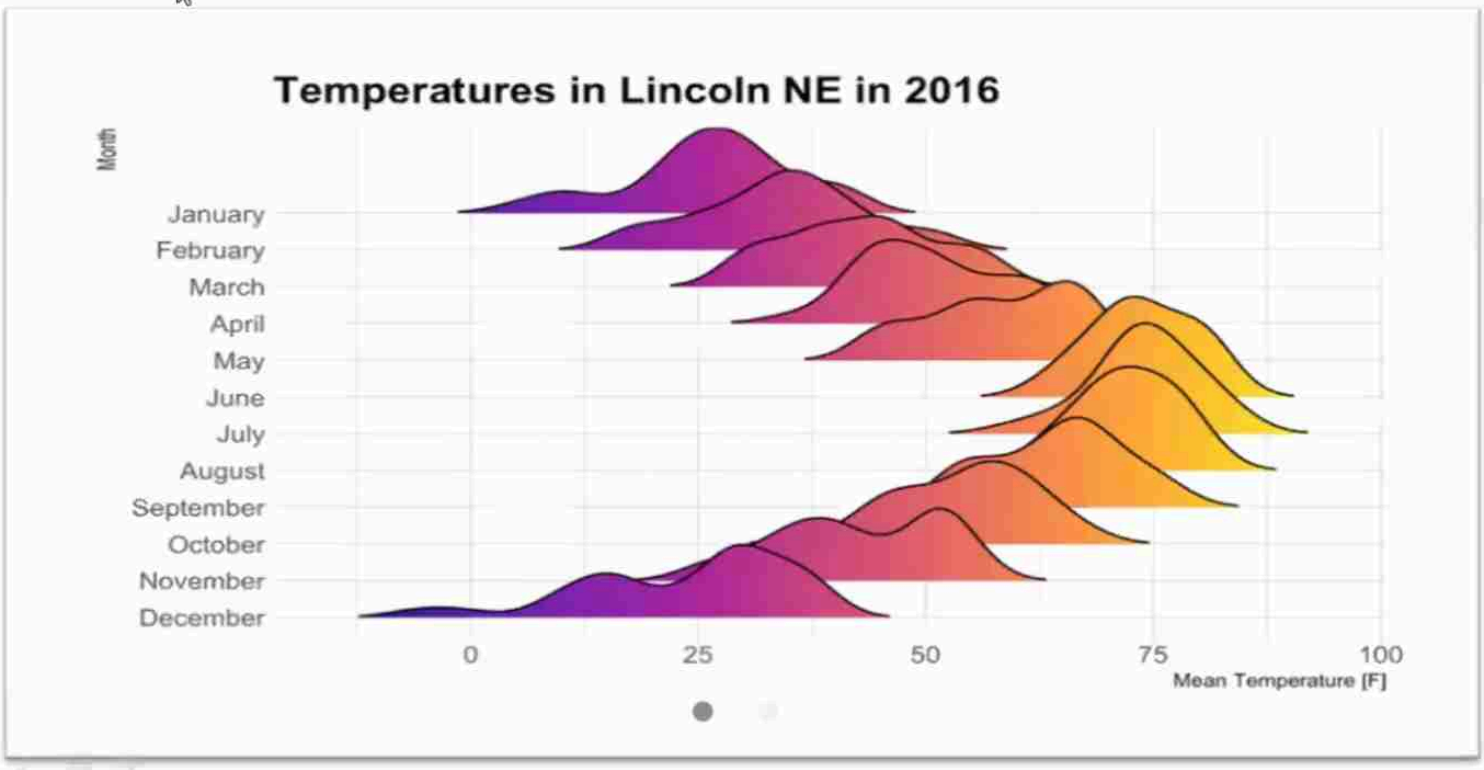
Variation

- The above example is a ridgeline plot using a set of density plots. It is possible to use histograms as well:



Cont'd...

- It is possible to color depending on the numeric variable instead of the categoric one.



Common mistakes

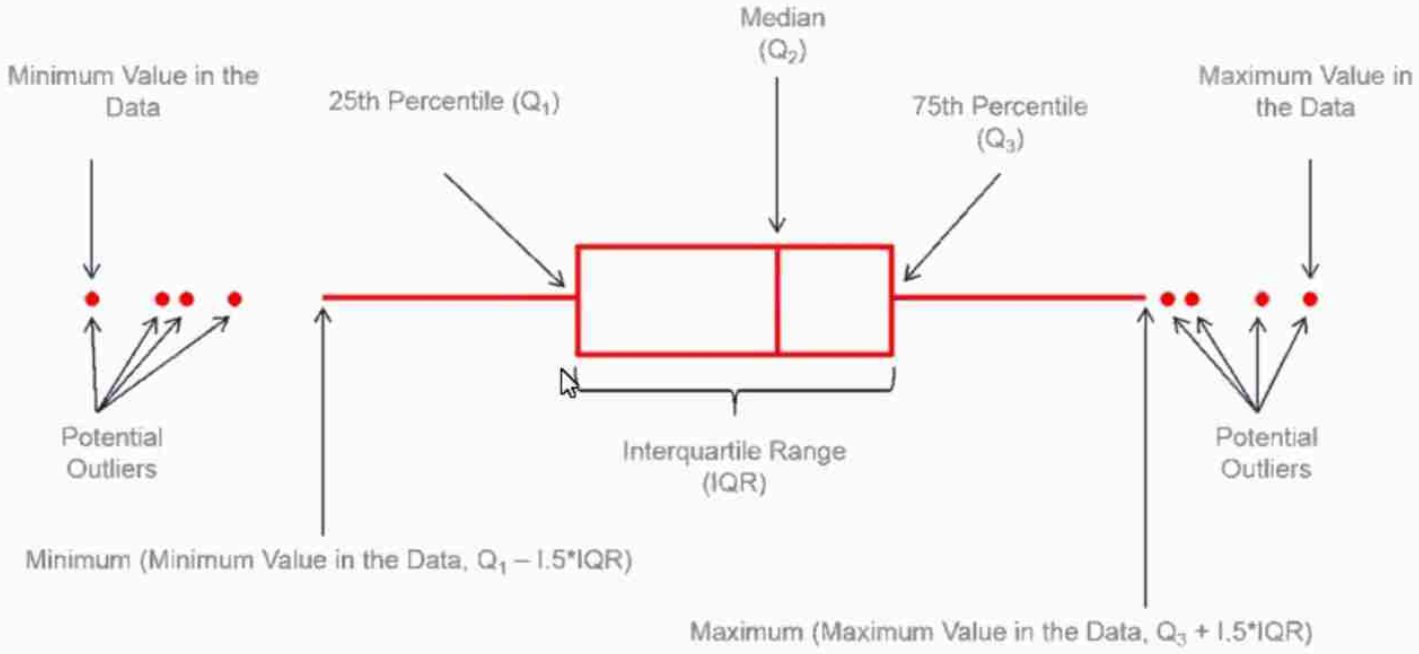
- As with histogram or density plot, play with bin size / bandwidth argument.
- Think about ordering groups in a smart way.
- Ridgeline plot works well when there is a clear pattern to discover since it hides a part of the data where the overlap takes place.

Boxplot

A boxplot gives a nice summary of one or more numeric variables. A boxplot is composed of several elements:

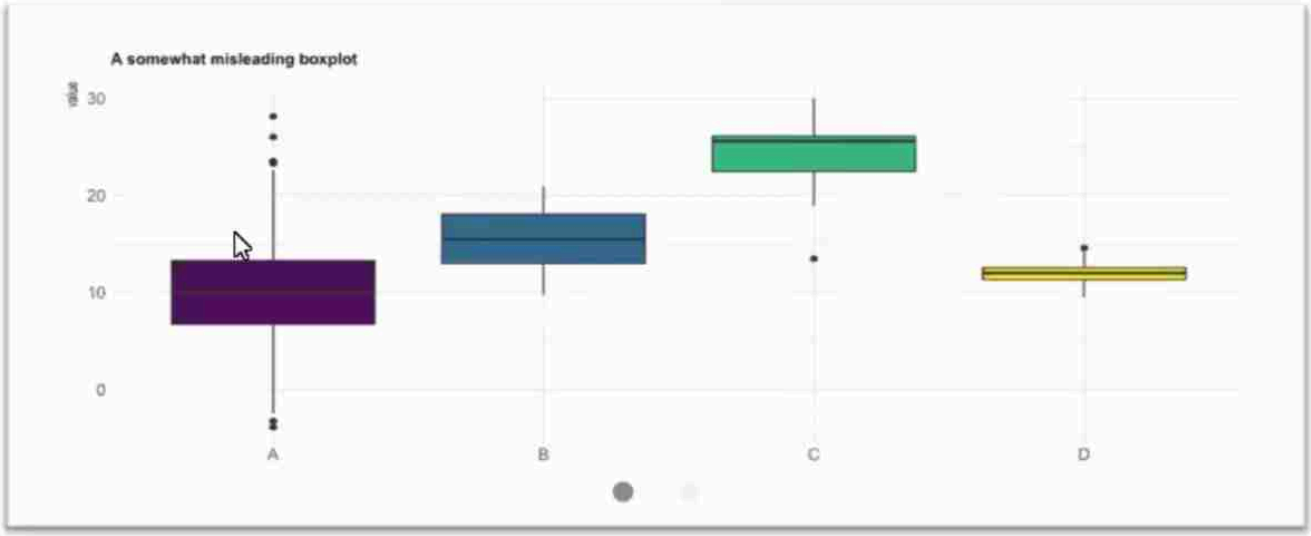
- The line that divides the box into 2 parts represents the median of the data. If the median is 10, it means that there are the same number of data points below and above 10.
- The ends of the box shows the upper (Q3) and lower (Q1) quartiles. If the third quartile is 15, it means that 75% of the observation are lower than 15.
- The difference between Quartiles 1 and 3 is called the interquartile range (IQR).
- The extreme line shows Q3+1.5x(IQR) to Q1-1.5x(IQR) (the highest and lowest value excluding outliers).
- Dots (or other markers) beyond the extreme line shows potential outliers.

Boxplot Anatomy



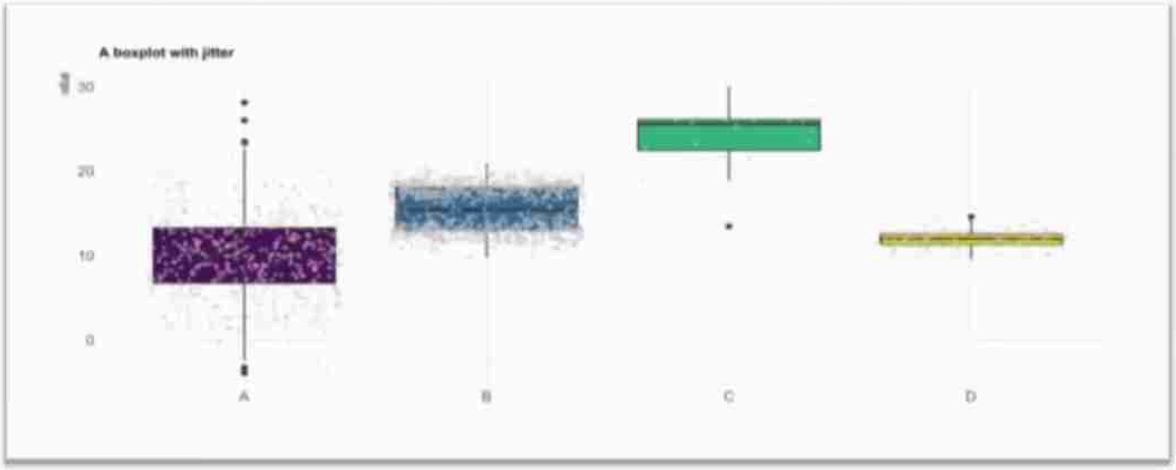
Cont'd...

- A boxplot can summarize the distribution of a numeric variable for several groups. The problem is that summarizing also means losing information, and that can be a pitfall. If we consider the boxplot below, it is easy to conclude that group C has a higher value than the others. However, we cannot see the underlying distribution of dots in each group or their number of observations.



Adding jitter

If the amount of data you are working with is not too large, adding jitter on top of your boxplot can make the graphic more insightful.



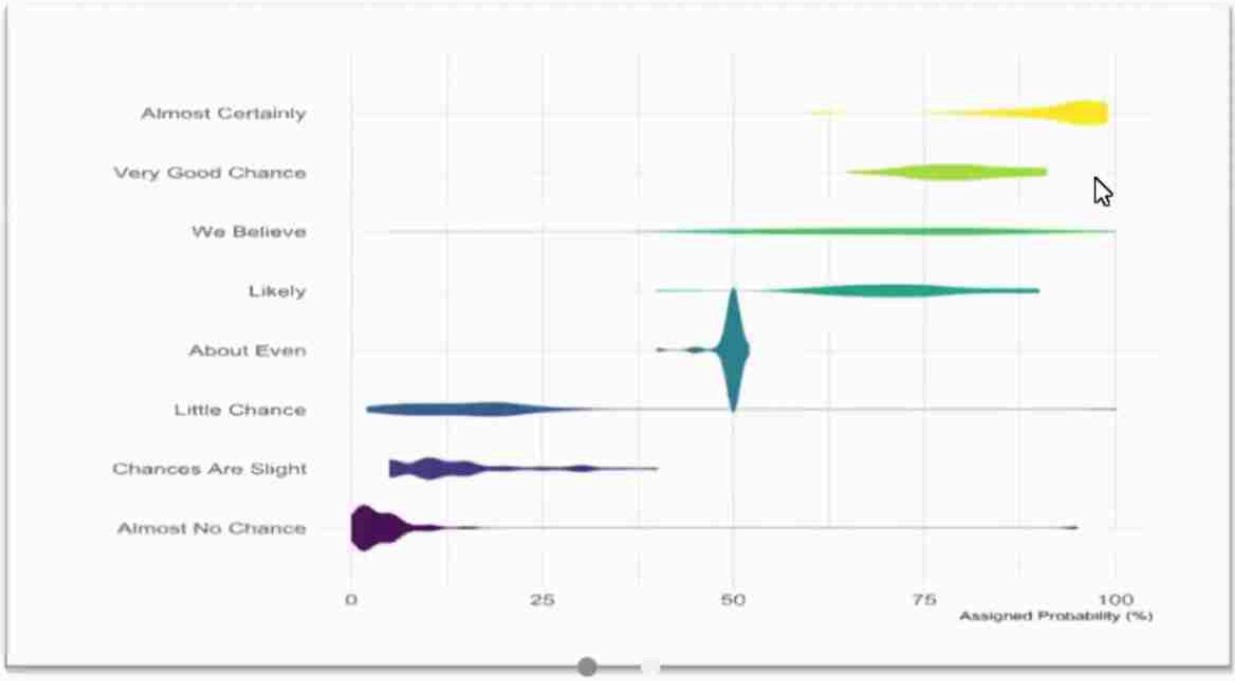
Here some new patterns appear clearly. Group C has a small sample size compared to the other groups. This is definitely something you want to find out before saying that group C has a higher value than the others. Moreover, it looks like group B has a bimodal distribution: dots are distributed in 2 groups: around y=18 and y=13.

Violin Plot

- Violin plot allows to visualize the distribution of a numeric variable for one or several groups. Each ‘violin’ represents a group or a variable. The shape represents the density estimate of the variable: the more data points in a specific range, the larger the violin is for that range. It is really close to a boxplot, but allows a deeper understanding of the distribution.

Example

- *What probability would you assign to the phrase "Highly likely"* were asked. Answers between 0 and 100 were recorded, and here is the distribution for each question:



een

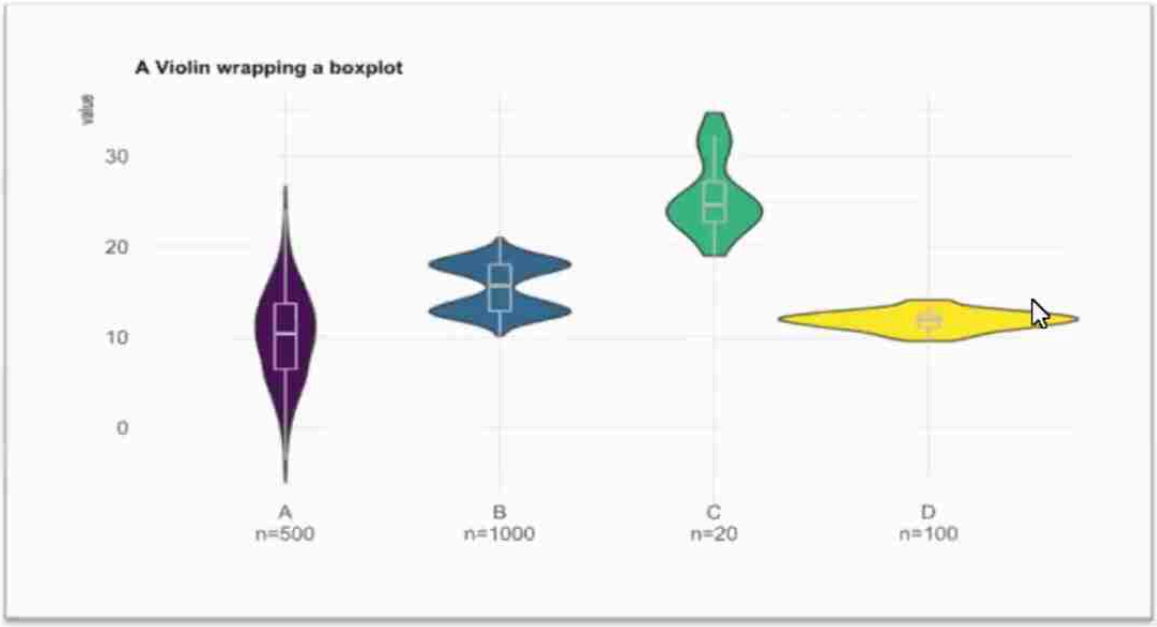
What for

- Violin plot is a powerful data visualization technique since it allows to compare both the ranking of several groups and their distribution. Surprisingly, it is less used than boxplot, even if it provides more information in my opinion.
- Violins are particularly adapted when the amount of data is huge and showing individual observations gets impossible. For small datasets, a boxplot with jitter is probably a better option since it really shows all the information.

Variation

- Violin plot are made vertically most of the time. If you have long labels, building a horizontal version like above make the labels more readable.
- It is possible to display a boxplot in the violin. It allows to assess the median and quartiles in a glimpse.

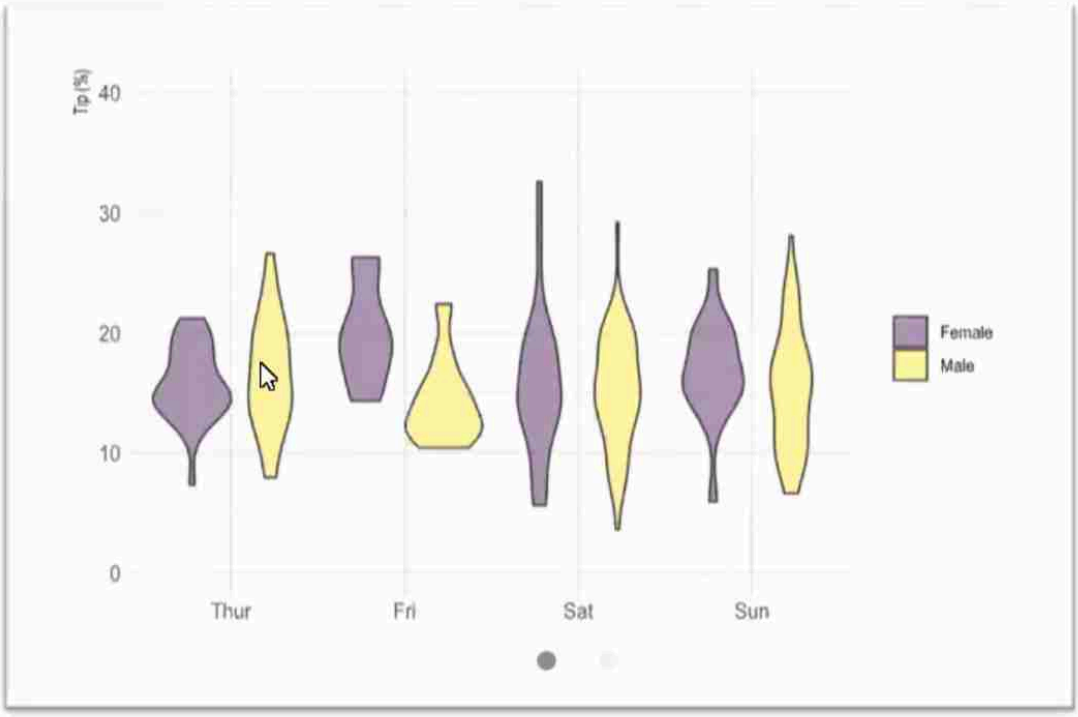
Cont'd...



- Here it is very clear that the groups have different distributions. The bimodal distribution of group B becomes obvious. Violin plots are a powerful way to display information–they are probably under-utilized compared to boxplots.

Cont'd...

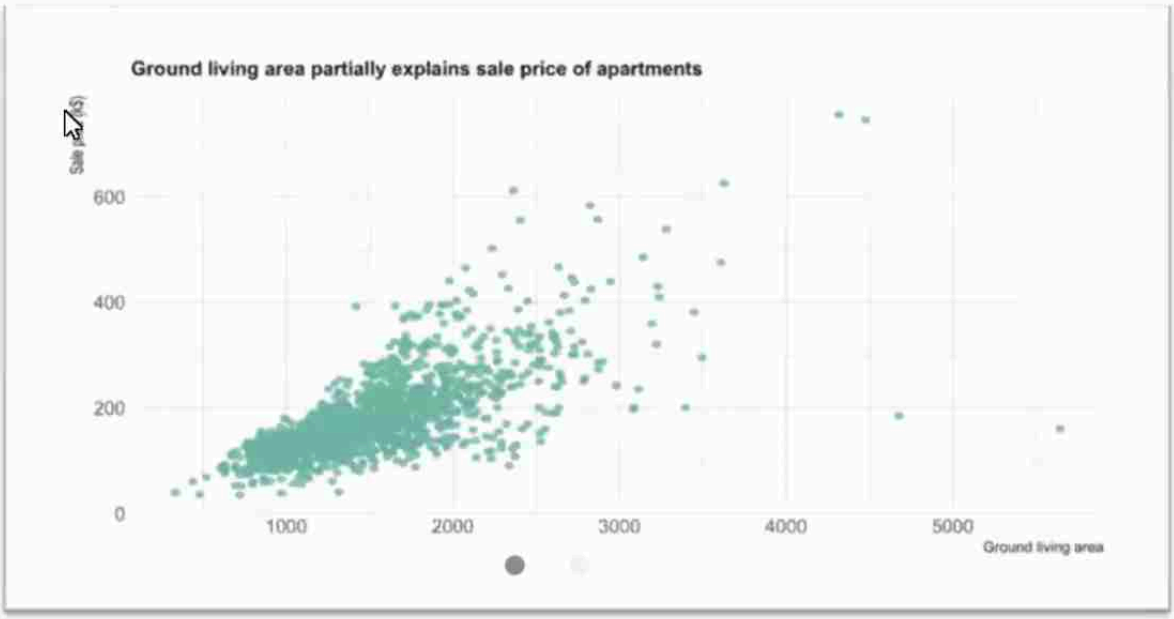
- If your variable are grouped, you can build a grouped violin as you would do for a boxplot. Here is an example showing how much Male and Female tip depending on the day of the week.



creen

Scatter Plot

- A scatterplot displays the relationship between 2 numeric variables. For each data point, the value of its first variable is represented on the X axis, the second on the Y axis.
- Here is an example considering the price of 1460 apartments and their ground living area.

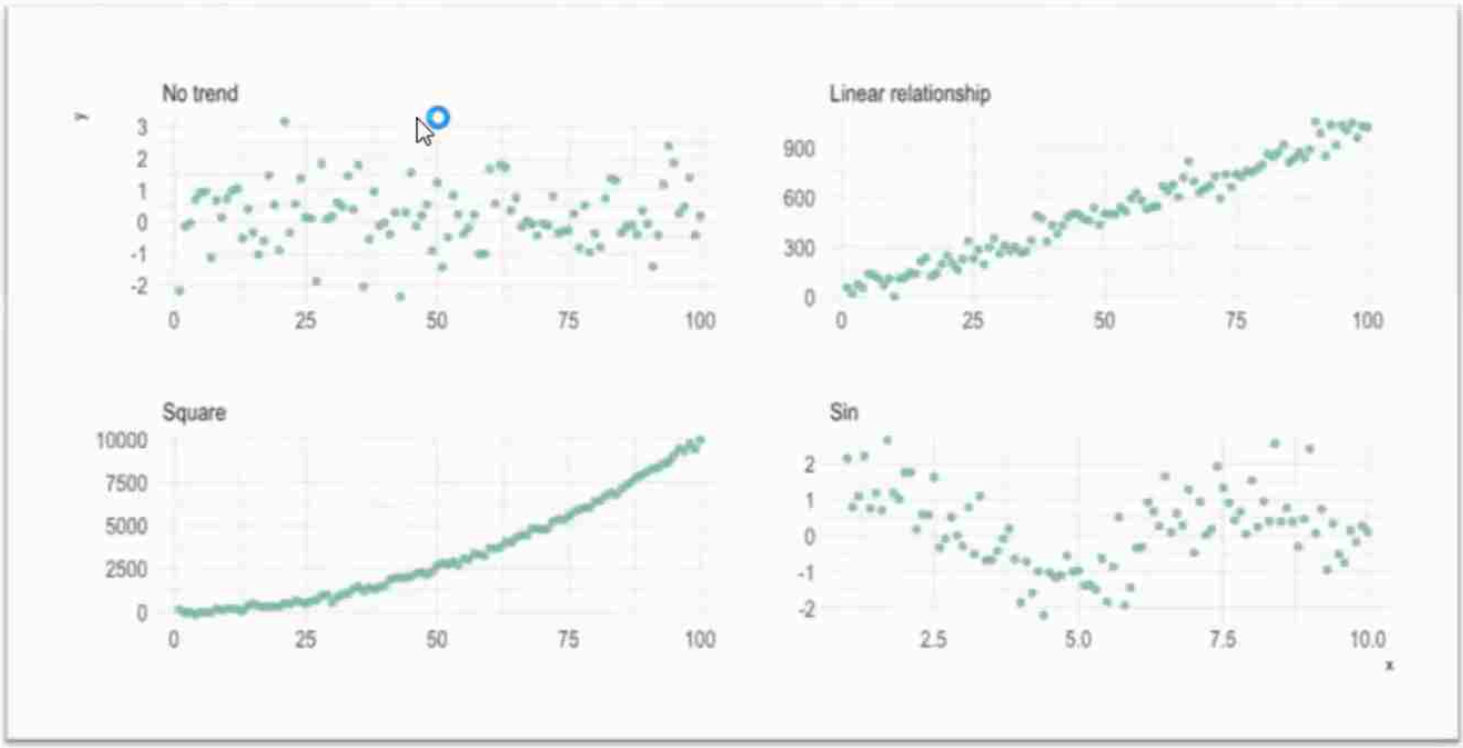


What for

- A scatterplot is made to study the relationship between 2 variables. Thus it is often accompanied by a correlation coefficient calculation, that usually tries to measure the linear relationship.
- However other types of relationship can be detected using scatterplots, and a common task consists to fit a model explaining Y in function of X. Here is a few pattern you can detect doing a scatterplot.

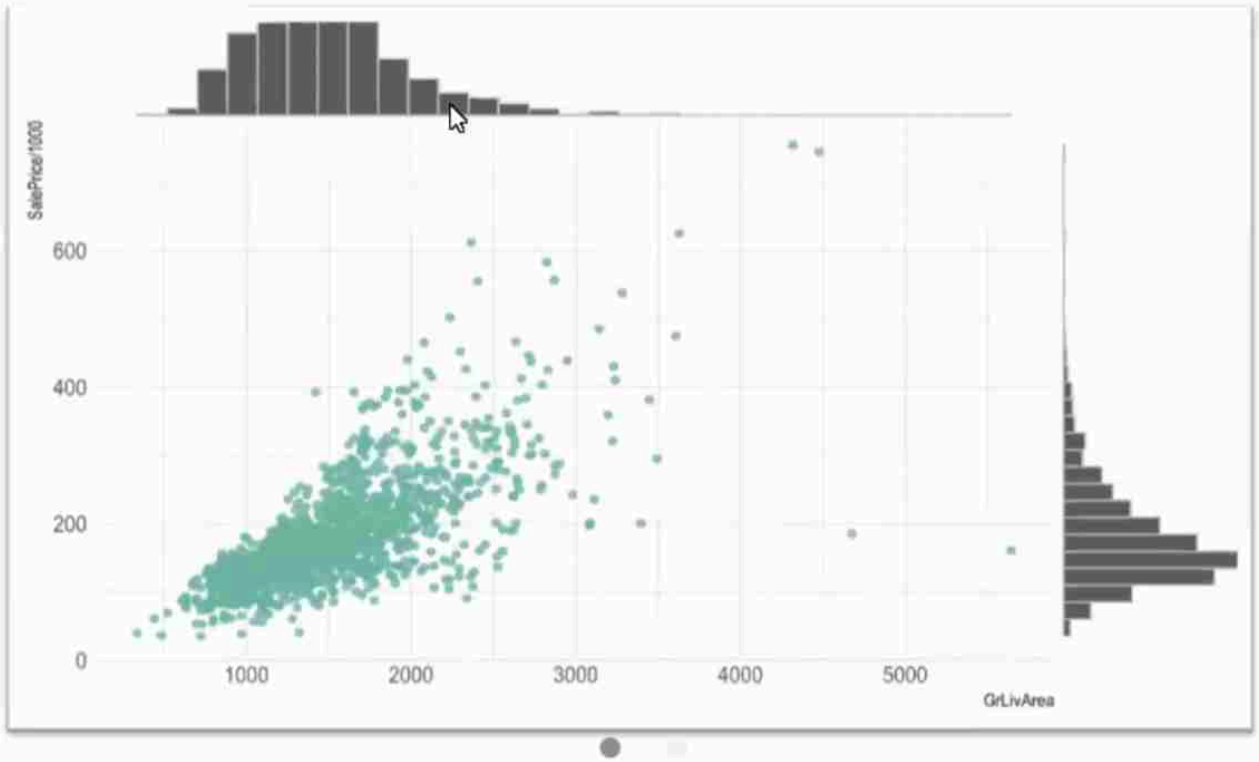
reen

Cont'd...


No trend
3
2
1
0
-1
-2
0
25
50
75
100
Linear relationship
900
600
300
0
0
25
50
75
100
Square
10000
7500
5000
2500
0
0
25
50
75
100
Sin
2
1
0
-1
-2
2.5
5.0
7.5
10.0
x

Cont'd...

- Scatterplot are sometimes supported by marginal distributions. It indeed adds insight to the graphic, revealing the distribution of both variables:

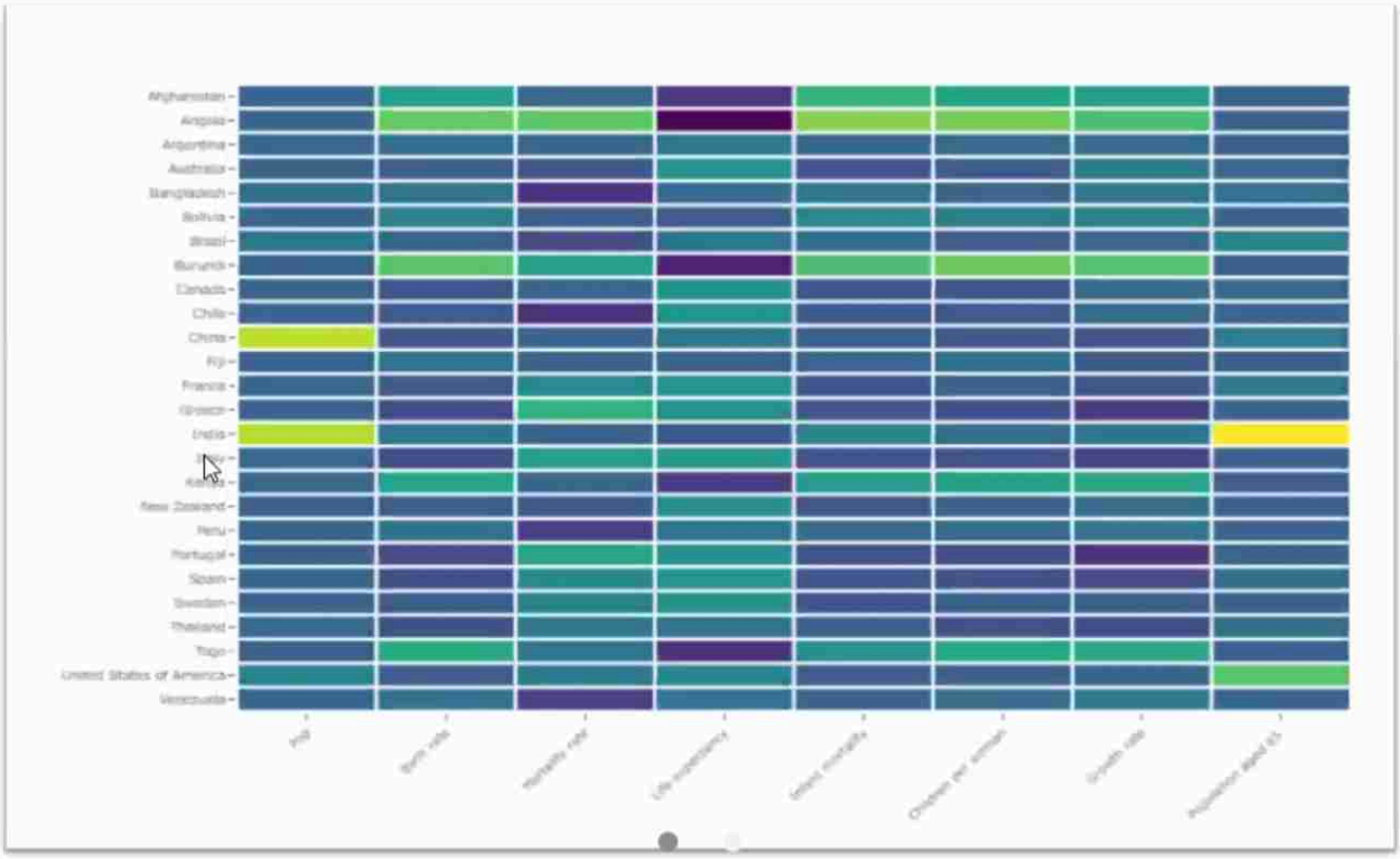


Heat Map

- A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is a bit like looking a data table.
- Here is an example showing 8 general features like population or life expectancy for about 30 countries in 2015.

Example

- Heatmap is really useful to display a general view of numerical data, not to extract specific data point.



Variation

- A common task is to compare the result with expectations. For instance, we can check if the countries are clustering according to their continent using a color bar.
- For static heatmap, a common practice is to display the exact value of each cell in numbers. Indeed, it is hard to translate a color in a precise number.
- Heatmaps can also be used for time series where there is a regular pattern in time.
- Heatmaps can be applied to adjacency matrix.

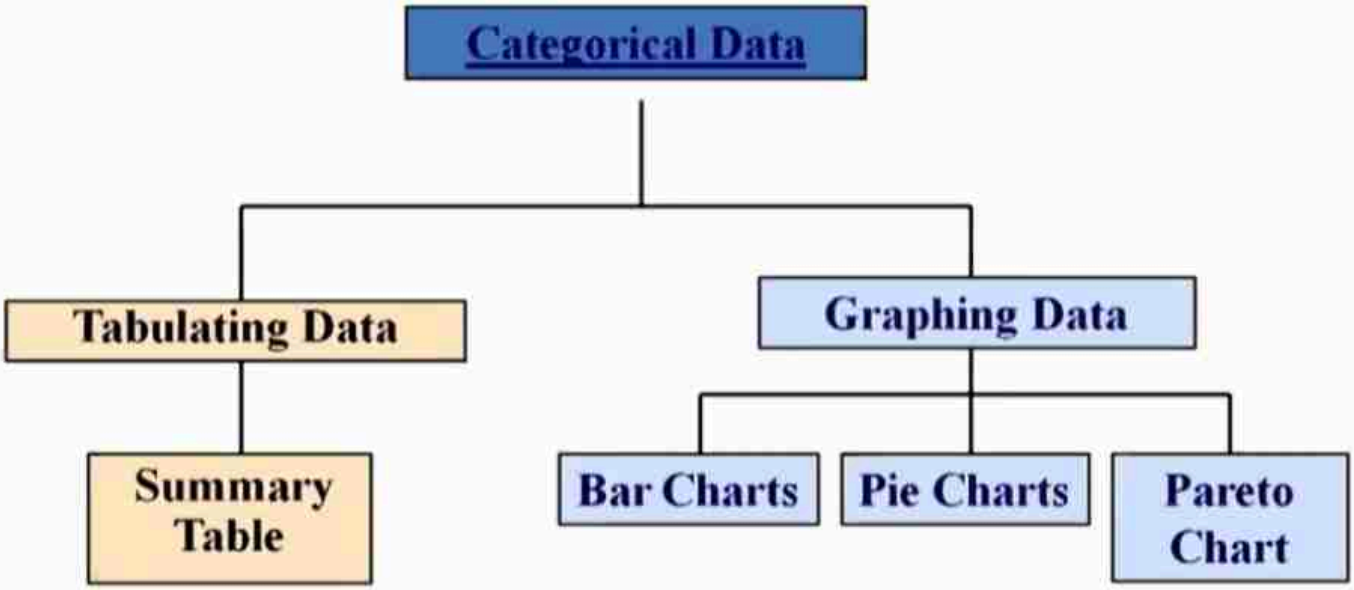
n

Common mistakes

- Often need to normalize your data.
- Use cluster analysis and thus permute the rows and the columns of the matrix to place similar values near each other according to the clustering.
- Color palette is important.

n

Categorical data are summarized by



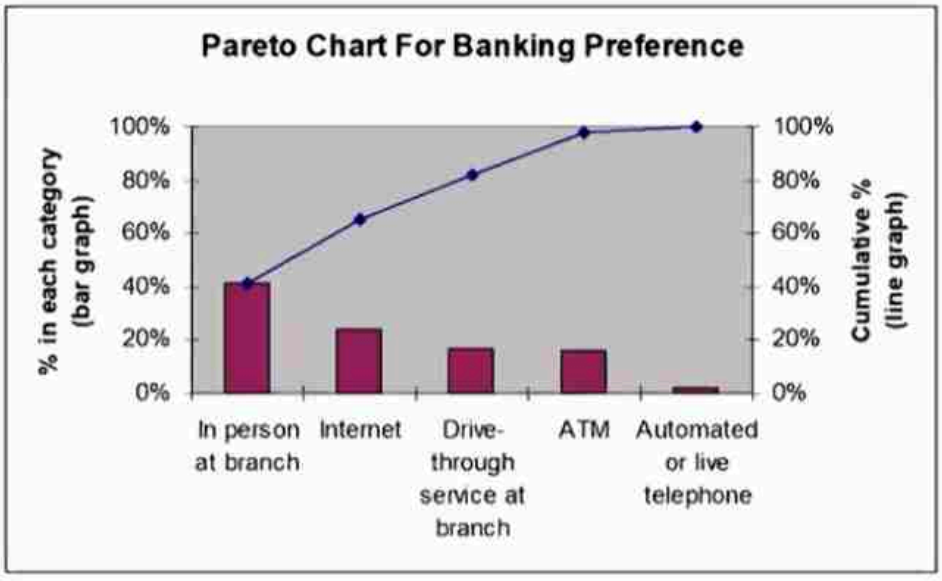
Summary Table

- A **summary table** indicates the frequency, amount, or percentage of items in a set of categories so that you can see differences between categories.

Banking Preference?	**Percent**
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

Pareto Chart

- Used to portray categorical data (nominal scale)
- **A vertical bar chart, where categories are shown in descending order of frequency**
- A **cumulative polygon** is shown in the same graph
- **Used to separate the "vital few" from the "trivial many"**



Ordered Array

- An **ordered array** is a sequence of data, in rank order, from the **smallest** value to the **largest** value.
- Shows **range** (minimum value to maximum value)
- May help identify **outliers** (unusual observations)
- Which values appear **more than one**
- Divide data in **sections** (Day students- 1/3rd of data below 18, 2/3rd below 22,etc)

Age of Surveyed College Students	**Day Students**					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Stem and leaf

- A simple way to see how the data are **distributed and where concentrations** of data exist

METHOD: Separate the sorted data series into **leading** digits (the **stems**) and the **trailing** digits (the **leaves**)

- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.

Age of Surveyed College Students	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Age of College Students

Day Students

Stem	Leaf
1	67788899
2	0012257
3	28
4	2

Night Students

Stem	Leaf
1	8899
2	0138
3	23
4	15

Cont'd...

Girls		Boys
7, 8, 2, 2, 1	1	5, 8
3, 3, 3, 2	2	2, 2, 3, 6
5, 4, 3	3	4, 5, 5, 5
7, 5, 4	4	0, 0, 2, 7, 9
1, 1, 0	5	0, 0, 1

Stems	Leaves
10	4 7
11	2 5 5 6
12	3
13	0 4
14	5 7

Means 145

8.	0	0						
9.	0							
10.	0	0						
11.	0	0	5					
12.	0	0	0	2				
13.	2	5	8	8				
14.	0	0	0	0	4	6	8	
15.	0	0	5					
16.	0	2	6	8				
17.	0	0	5					
18.	0	2	5					
19.	0	5						
20.	0	5						

Decimal Between Stem and Leaf

12.3, 12.5, 13.0

Becomes

12	3, 5
13	0

Key: 12 | 3 = 12.3 units

Decimal in the Stem

1.23, 1.25, 1.30

Becomes

1.2	3, 5
1.3	0

Key: 1.2 | 3 = 1.23 units

Frequency Distribution

- The **frequency distribution** is a summary table in which **the data are arranged into numerically ordered classes.**
- You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.
- The number of classes depends on the number of values in the data. With a **larger** number of values, typically there are **more classes**. In general, a frequency distribution should have at **least 5 but no more than 15 classes**.
- To determine the **width of a class interval,** you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

Example

A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: **58 - 12 = 46**
- Select number of classes: **5** (usually between 5 and 15)
- Compute class interval (width): **10** (46/5 then round up)
- Determine class boundaries (limits):
  - **Class 1: 10 to less than 20**
  - **Class 2: 20 to less than 30**
  - **Class 3: 30 to less than 40**
  - **Class 4: 40 to less than 50**
  - **Class 5: 50 to less than 60**
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes

Cont'd...

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Cumulative Frequency

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

Why do we use Frequency Distribution?

- It **condenses** the raw data into a more useful form
- It allows for a quick **visual interpretation** of the data
- It enables the determination of the major characteristics of the data set including **where the data are concentrated / clustered**

Some Tips

- Different **class boundaries** may provide **different pictures** for the same data (especially for smaller data sets)
- **Shifts in data concentration** may show up when **different class** boundaries are chosen
- As the **size of the data set increases**, the impact of alterations in the **selection of class boundaries is greatly reduced**
- When comparing two or more groups with **different sample sizes,** you must use either a **relative frequency or a percentage distribution**

Polygon

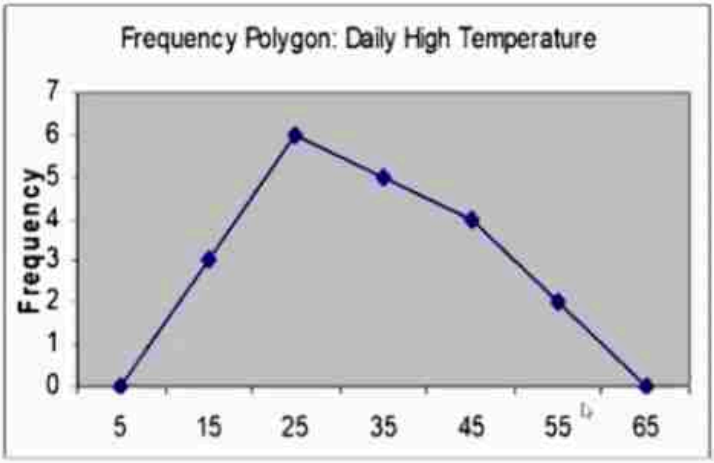
- A **percentage polygon** is formed by having the **midpoint of each class represent the data in that class and then connecting the sequence of midpoints** at their respective class percentages.

- The **cumulative percentage polygon,** or **ogive,** displays the variable of interest along the *X* axis, and the cumulative percentages along the *Y* axis.

- **Useful when there are two or more groups to compare**.

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

(In a percentage polygon the **vertical axis** would be defined to show the **percentage of observations per class**)



Ogive (Cumulative % Polygon)

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Class	Lower class boundary	% less than lower boundary
10 but less than 20	10	15
20 but less than 30	20	45
30 but less than 40	30	70
40 but less than 50	40	90
50 but less than 60	50	100

(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.