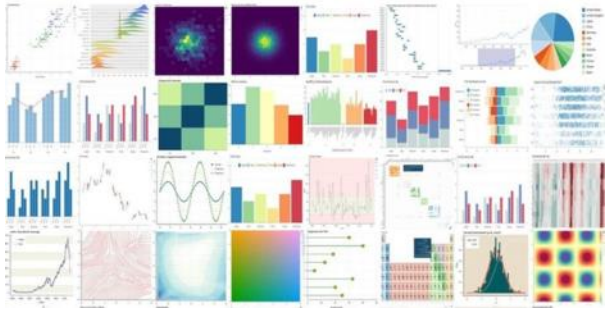
Data Visualization & Business Intelligence with Python

Making data more understandable



Muhammad Adeel Sultan Khan

Data Scientist Agtech, Silicon Valley, California

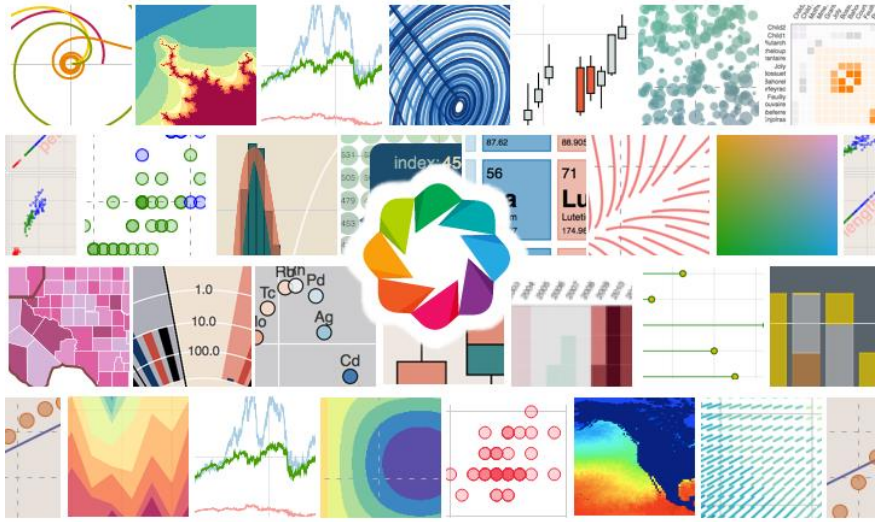
Personal Background

- Muhammad Adeel Sultan Khan
- Ag tech Data Scientist Silicon Valley tech company over 5 years
- Treasury Finance 8 years experience
- MSIS Data Science Analytics, Santa Clara University
- MBA San Jose State University
- CFA Level 1 exam
- BCS & MBA MIS, Institute of Business Administration

Agenda

- **Segment 1: Data Visualization Foundations**
    - What is data visualization
    - Explore the conceptual and technical fundamentals of data visualization
    - Why is data visualization important
    - What are the advantages/disadvantages of data visualization
    - Why is data visualization important for big data
    - What are the different types of visualizations
    - How to create persuasive data visualizations for audience
    - What is the claim-warrant framework in creating effective visualizations

- **Segment 2: Introduction to Business Intelligence**
    - What is business intelligence
    - How business intelligence works
    - Different types of BI methods
    - How BI, data analytics, and business analytics work together
    - Traditional BI vs. modern BI
    - What are the benefits of business intelligence
    - The categories of BI analysis: predictive, descriptive, & prescriptive
    - Explore BI dashboards and why are they important
    - Popular BI tools
    - Data wrangling with Python



Agenda

- **Segment 3: Data Visualization & Business Intelligence in action**
  - Data Visualization Redesign – LinkedIn top skills dataset
  - Deceptive Visualization – Australia's Gender Pay Gap
  - Should you be tempted to invest in cryptocurrencies – Using data viz to support the claim
  - Exercise/Activity Title: Demo of data wrangling and creation of data viz in a python jupyter notebook

What is Data Visualization

- Data visualization is a graphical representation of information & data.
- By using visual elements like charts, graphs & maps, data visualization tools provide an accessible way to find trends, outliers, and patterns in data. Visualizations help to present data to non-technical audiences without confusion.
- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Why data visualization is important

- Data Visualization helps people to see and understand data better
- An effective visualization brings people on same page regardless of their expertise
- Every field government, finance, marketing, healthcare, consumer goods, sports is benefiting from making data more understandable
- Data science skills are in high demand & having good data visual skills is imperative
- Traditional education draws a distinct line between creative storytelling and technical analysis, the modern professional world also value those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling

Advantages of data visualization

- Data collection rate is increasing exponentially and visualizations help in understanding the data . Some advantages of data visualization include:

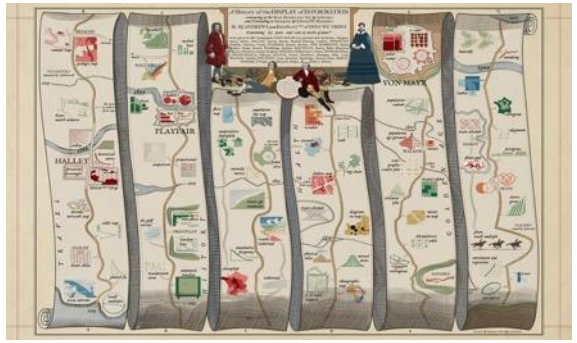
  - ❖ Show trends, patterns & outliers. Display complex relationships
  - ❖ Color patterns differentiates between important attributes
  - ❖ Share information/data easily with the intended audience
  - ❖ Enable teams to take appropriate actions or strategic decision
  - ❖ Unlocking key values for strategic decision making in business

Disadvantages of data visualization

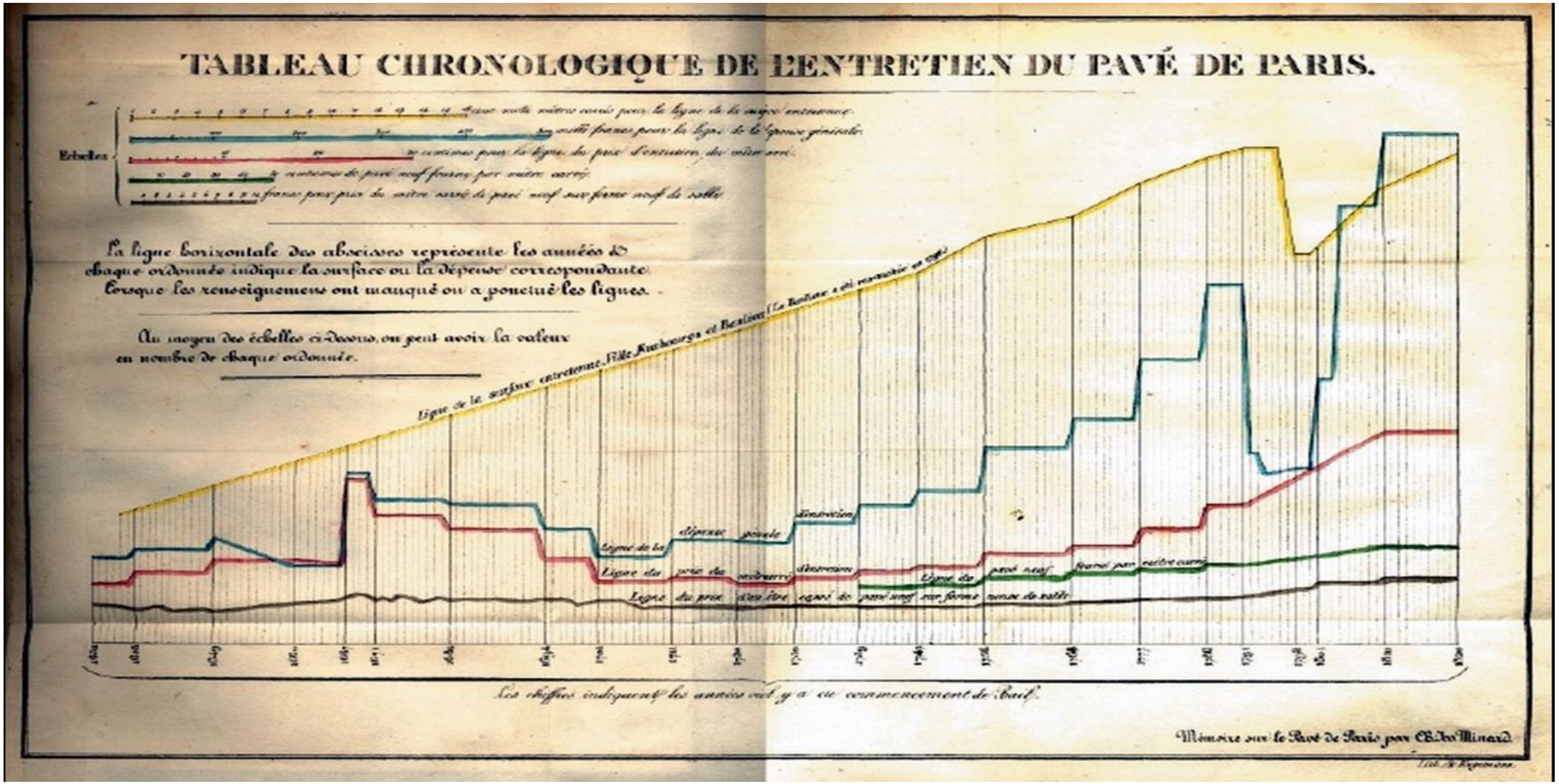
- When viewing a visualization with many different datapoints, it's easy to make an inaccurate assumption. Or sometimes the visualization is just designed wrong so that it's biased or confusing.
- Some other disadvantages include:
  - ❖ Biased or inaccurate information.
  - ❖ Correlation doesn't always mean causation.
  - ❖ Core messages can get lost in translation.
  - ❖ Data wrangling not done properly could lead to inaccuracy

HISTORY OF DATA VISUALIZATION

- **Charles Joseph Minard:** (27 March 1781 – 24 October 1870) was a French civil engineer recognized for his significant contribution in the field of information graphics in civil engineering and statistics. Minard was, among other things, noted for his representation of numerical data on geographic maps especially his flow maps.
- Minard's first statistical graphic, from 1825, depicts several time series related to Paris pavement maintenance over the preceding two centuries.
- Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates. Statistician professor Edward Tufte described the graphic as what ***"may well be the best statistical graphic ever drawn".***

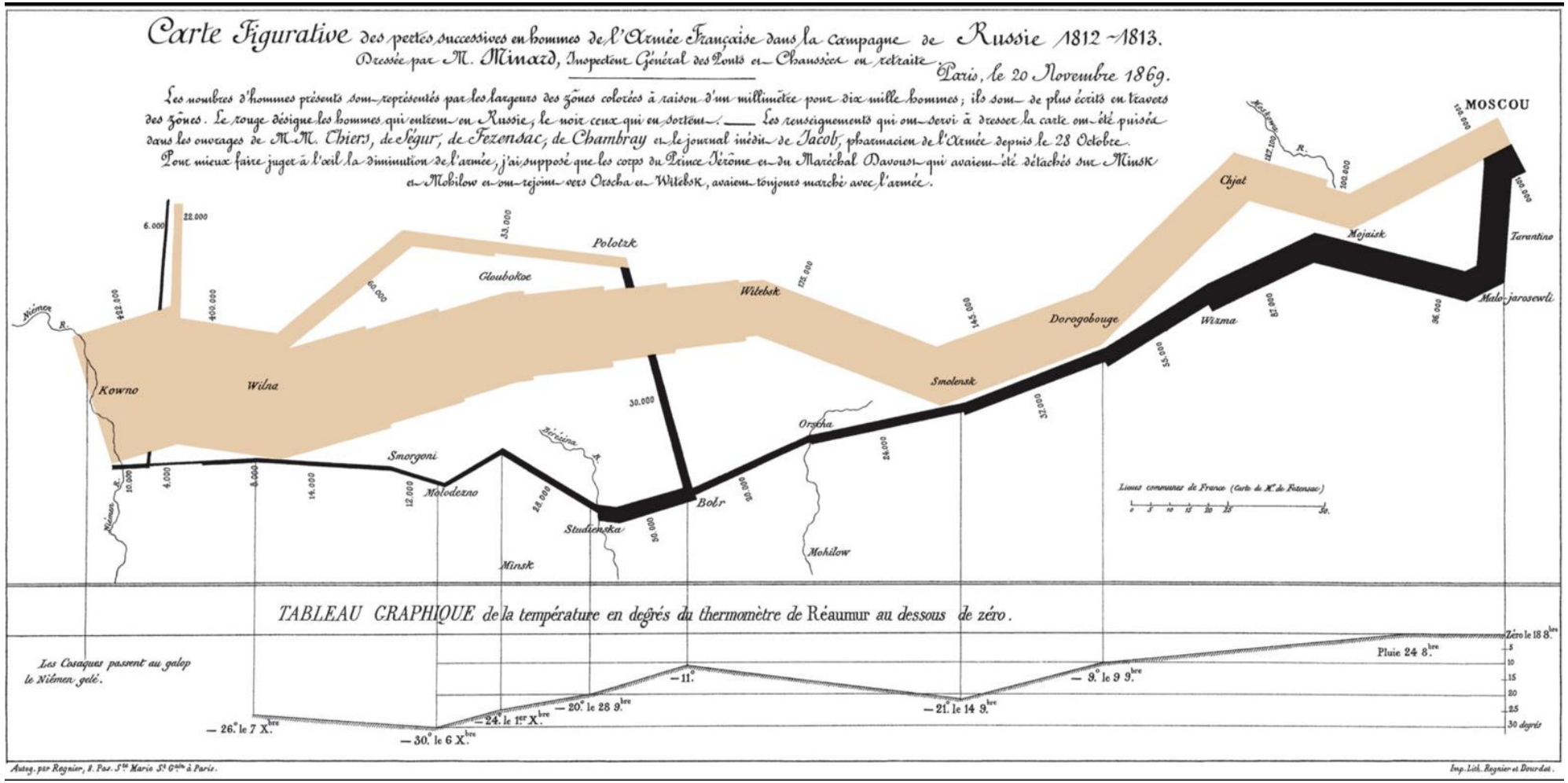


HISTORY OF DATA VISUALIZATION



Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates.

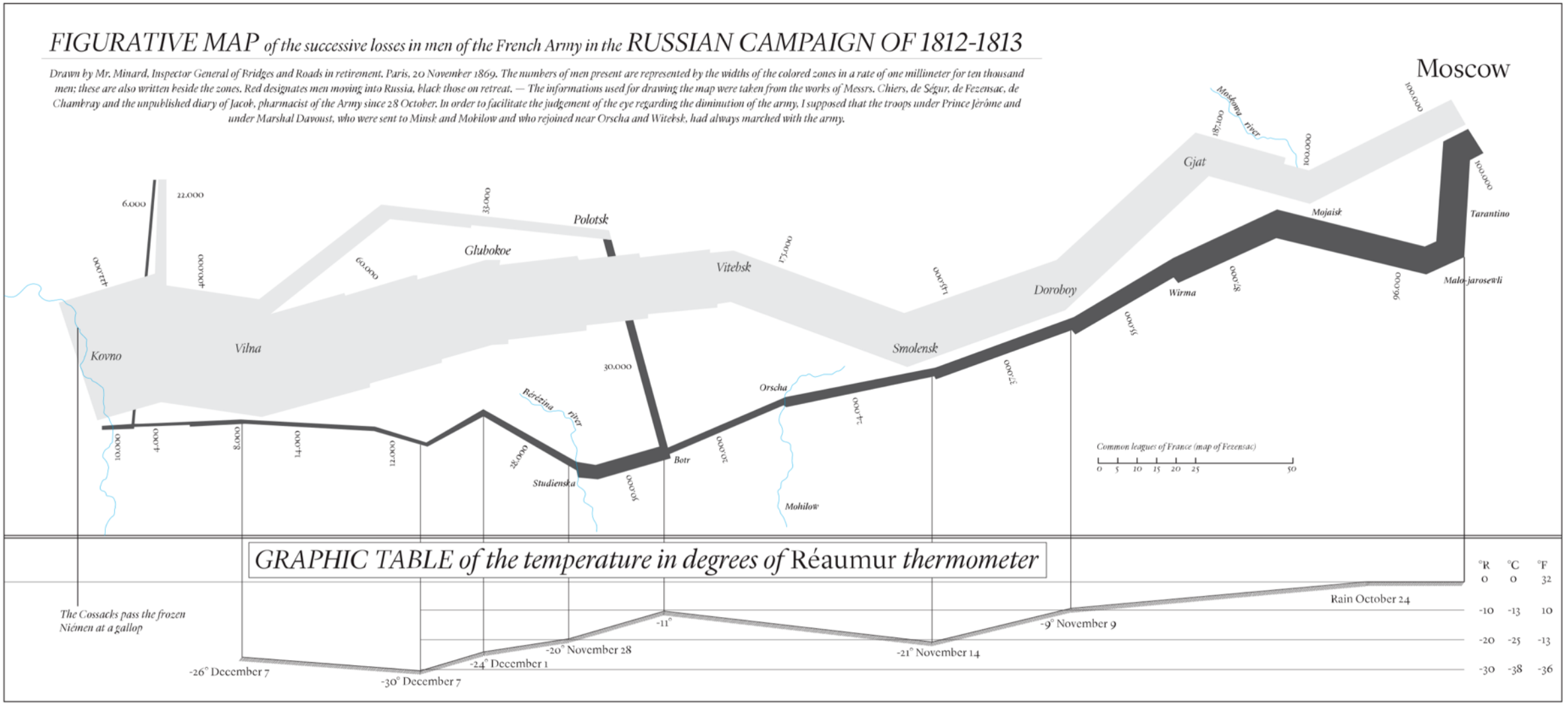
HISTORY OF DATA VISUALIZATION



tan represents the men who invaded Russia itself, while the black represents the retreat from Moscow. The widths of the colored zones shrink serves as a sobering representation of the staggering numbers of men who were lost in the course of the six-month campaign.

The orange and black columns crossing the map show the French *Grande Armée* on its march to and from Moscow. The width of the column shows the size of the army – ever-shrinking as thousands of Napoleon's soldiers died of disease, cold, starvation and Russian attacks. This image brilliantly displays the devastation of Napoleon's army over the winter of 1812-13. It has been called "the best statistical graph ever drawn".

REDRAWING OF CHARLE'S 1812 NAPOLEAN'S CHART

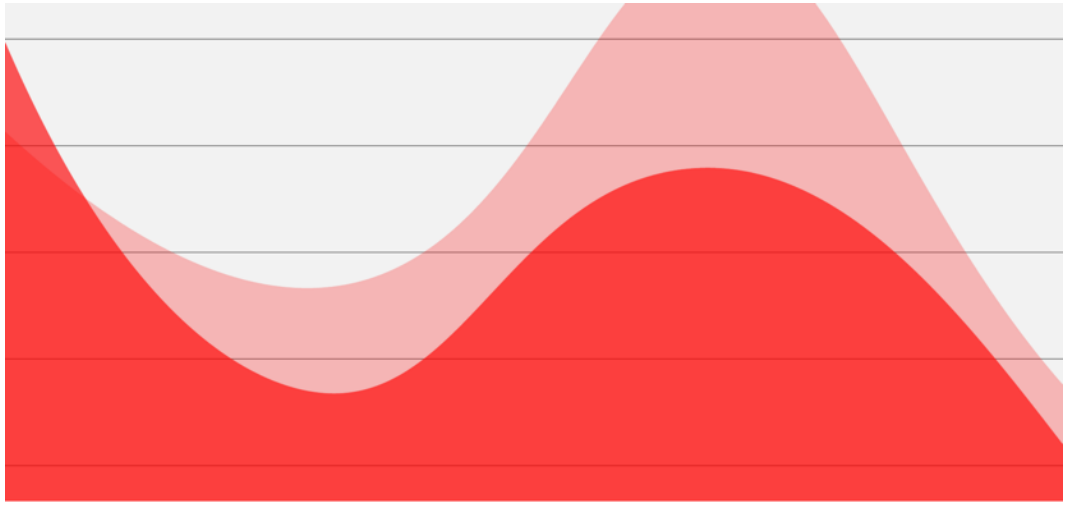


Modern redrawing of Charles Joseph Minard's figurative map of the 1812 French invasion of Russia, including a table of temperatures converting degrees Réaumur to degrees Fahrenheit and Celsius

Why Data Visualization

- Data visualization can be utilized for a variety of purposes, and it's not only reserved for use by data teams
- Management also leverages it to convey organizational structure and hierarchy while data analysts and data scientists use it to discover and explain patterns and trends
- Harvard Business Review categorizes data visualization into ***4 key purposes*** :

  - Idea Generation
  - Idea Illustration
  - Visual Discovery
  - Everyday dataviz



Why Data Visualization

- **Idea Generation:**
  - Data Visualization is used to generate ideas in teams using identified trends, patterns and relationship from the visualizations.
  - Data Visualizations are used in discussion and brainstorming sessions by supporting different perspectives and highlighting the common concerns or viewpoints
  - Data Visualizations align teams and whole company on common goals and strategic objectives by magnifying the real scenarios in easy to understand terms

Why Data Visualization

- **<u>Idea Illustration:</u>**
  - Data Visualizations for idea illustration assists in conveying an idea, such as a tactic or process
  - It is commonly used in learning settings, such as tutorials, certification courses, centers of excellence, but it can also be used to represent organization structures or processes, facilitating communication between the right individuals for specific tasks
  - Project managers frequently use ***Gantt charts*** and ***Waterfall charts*** to illustrate workflows
  - Data Modeling also uses abstraction to represent data flow within an enterprise's information system, making it easier for developers, business analysts, data architects to understand the data relationships.

Why Data Visualization

- **Visual Discovery:**

  - Visual discovery and every day data viz are more closely aligned with data teams. While visual discovery helps data analysts, data scientists, and other data professionals identify patterns and trends within a dataset, every day data viz supports the subsequent storytelling after a new insight has been found.

What is data visualization good for

- Visualizations are tools that can make complex concepts easier for humans to understand.
- In the words of engineer and inventor Douglas Engelbart, “a tool doesn’t just make something easier—it allows for new, previously-impossible ways of thinking, of living, of being.”
- The utility of data visualization can be divided into three main goals: ***to explore, to monitor, and to explain***. While some visualizations can span more than one of these, most focus on a single goal.

What is data visualization good for

- **To Explore:**

  - When users are looking for an open-ended tool that helps them to find patterns and insights in data, a data visualization focused on exploration and fast iteration can help.

  - Exploration tools should have strong connections to other tools that collect (extract), clean (transform), and curate (load) data.

What is data visualization good for

- **To Monitor:**

  - When users need to check on performance, a data visualization focused on monitoring is best.
  - Monitoring tools, such as dashboards, should focus on leading indicators and showing information that is connected to useful and direct actions.
  - Leading indicators predict future performance and help drive your daily initiatives.
  - Lagging indicators reflect past performance to assess success and shape long-term strategy.

What is data visualization good for

- **To Explain:**

  - When users want to go beyond the “what” of a problem and dig into the “why,” a data visualization focused on explanation is ideal.

  - Explanatory visualizations are often hand-crafted to help a broad audience understand a complex subject, and usually are not able to be automated.

Why use data visualization

- According to IBM, 2.5 quintillion bytes of data are created every day. The Research Scientist Andrew McAfee and Professor Erik Brynjolfsson of MIT point out that "**more data cross the internet every second than were stored in the entire internet just 20 years ago.**"
- As the world becomes more connected with increasing number of electronic devices, volume of data will continue to grow exponentially. IDC predicts there will be 175 zettabytes of data by 2025.
- Raw data is hard for human brain to comprehend— it's difficult for human brain to comprehend numbers larger than five without drawing some kind of analogy or abstraction.
- Big data is useless without understanding. That's why data visualization plays an important role in everything from economics to science and technology, to healthcare and human services.
- By turning raw data into graphs, content becomes easier to understand and use.

Data visualization principles

- **Define a Clear Purpose:**
  - Before creating a data viz, defining a clear purpose will make the process easier and useful and prevent wasting time creating visuals that are not required.
  - Data visualization helps in strategic decision making and providing answers to hypothesis. It can be used to track performance, monitor customer behavior, and measure effectiveness of processes.
- **Know the Audience:**
  - Identifying audience before creating data viz is absolute must as it should be compatible with audience expertise and help viewers to view and process data easily and quickly.
  - Audience preferences guides every decision about visualization – dissemination mode, graph type, formatting & color. Audience intellectual level is an important feature to identify before designing.

Data visualization principles

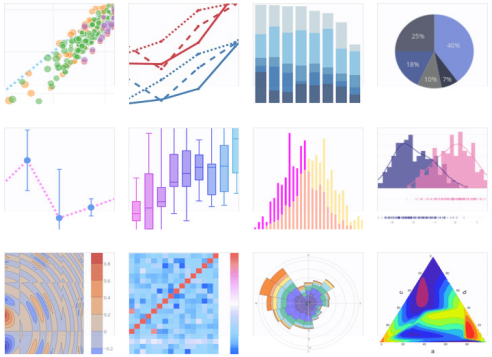
- **<u>Know the Audience:</u>**
  - We should identify what information audience already has and what additional information can our charts provide and focusing mostly on that missing info. to include in the charts
  - Some chart types are not suitable for a certain group of audience so it ‘s important to match charts with audience needs, requirements and expertise.
  - Before designing a data visualization, we must answer the following:
    - Who will be using this data?
    - What are their objectives?
    - How will they interact with the data?
    - What business questions do users need answered?

Data visualization principles

- **Keep it simple:**
  - Simplicity allows us to visually derive conclusions from data more easily. Ineffective visualizations (such as lengthy tables or complex charts) require more conscious thought to analyze information — slowing viewers down and reducing impact.
  - To quote Stephen Few, one of the founders of modern data visualization, “A picture is worth a thousand words...but only when the story is best told graphically rather than verbally and the picture is well designed.”
- **Use colors wisely:**
  - Color is a great way to focus your viewer’s attention and it’s best not to overwhelm people with too much of it.
  - When using color in data visualizations, you need to think about both hue (what color something is) and intensity (how saturated the color is.)
  - You should also keep your use of color consistent. For instance, only use a new color when it relates to a different meaning in the data.

Data visualization principles

- **<u>Use the right chart type:</u>**
  - The primary purpose of data visualization is to help viewers discover actionable insights. When selecting a chart type, you need to ask yourself:
    - What story you're trying to tell
    - Who your data visualization is for, and what their priorities are
    - What problem you're trying to solve
    - What type of data you have (for instance, do you have categories of data? If so, how many categories?)



Data visualization principles

Use the right chart type:

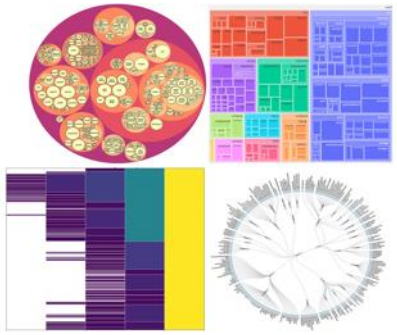
- **Pie charts** are best for comparisons between a relatively low number of categories;
- **Bar charts** are best for precise comparisons between categories, and for when you want to show negative and positive values in the dataset;
- **Scatter charts** are great to show correlation and clustering, especially if you have a lot of data
- **Line charts** emphasize trends over time;
- **Bubble charts** showcase distribution or relationships in large data sets;
- **Area charts** let you compare volumes of data easily.

Data visualization principles

- **Hierarchy:**
  - An important principle in data visualization and dashboard design is hierarchy. There is an expected widget arrangement that users will understand intuitively, without any extra explanation. The right hierarchy is what helps make visualized data and dashboards easily scannable.

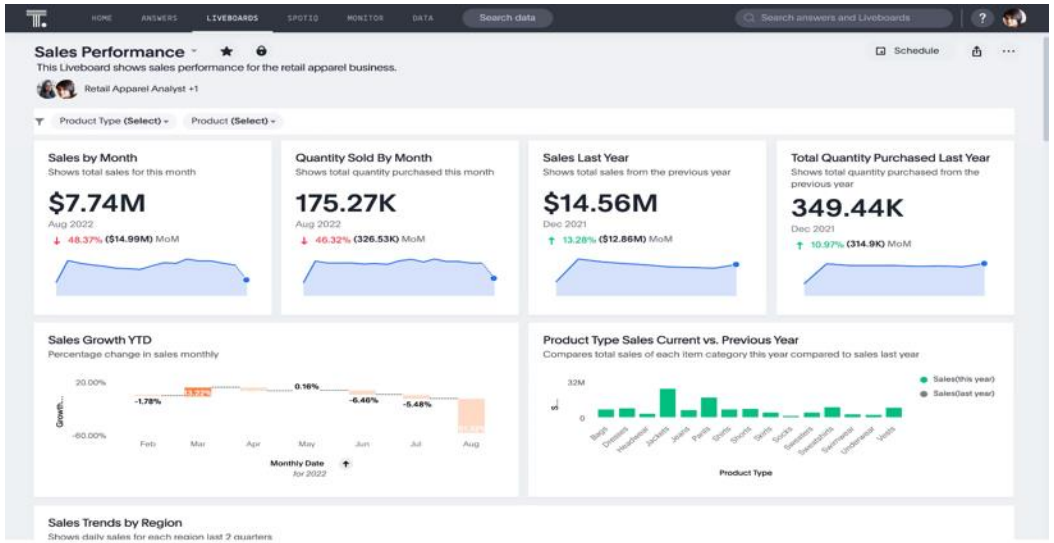
    Hierarchy has a few simple concepts:

    - The most important information should be located at the top left corner of the dashboard as this is a prime location and it's the first place that a viewer looks.
    - Widgets that follow the top-level should explain in more detail a top-level key performance indicator
    - Text widgets can be used to create titles that can help the content be more clear.
    - White space is important it's better to leave a gap rather than make something bigger just to fill space.
  - Visual hierarchy is crucial in data viz as it helps viewers quickly get the important info. Organizing information in a logical order and grouping related elements together can contribute to better visual hierarchy.



Data visualization principles

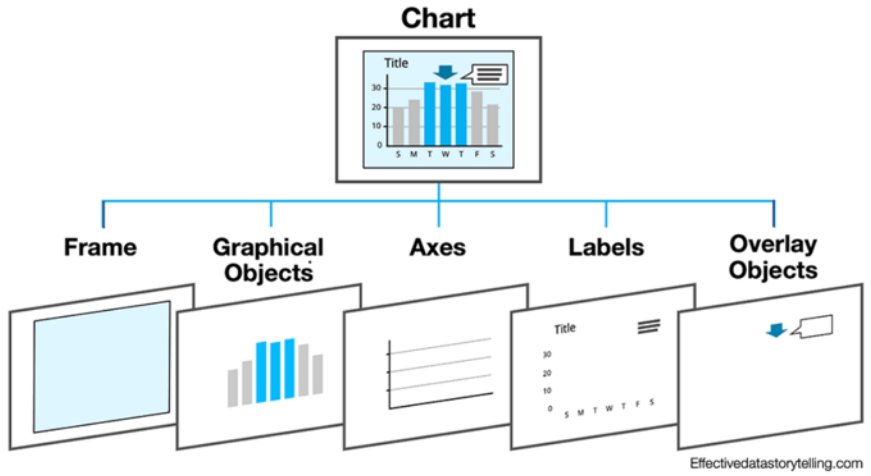
- **<u>Highlight the important information:</u>**
  - Data visualizations tell a story. By highlighting the most crucial information first, viewers can see what the story is all about.
  - You can also use the position of the data to create emphasis. For Western audiences, the top-left is considered prime real estate.
  - e.g., total sales by month is likely the most important information for audience, Liveboard show that data in the top-left.



https://www.thoughtspot.com/data-trends/data-visualization/data-visualization-principles

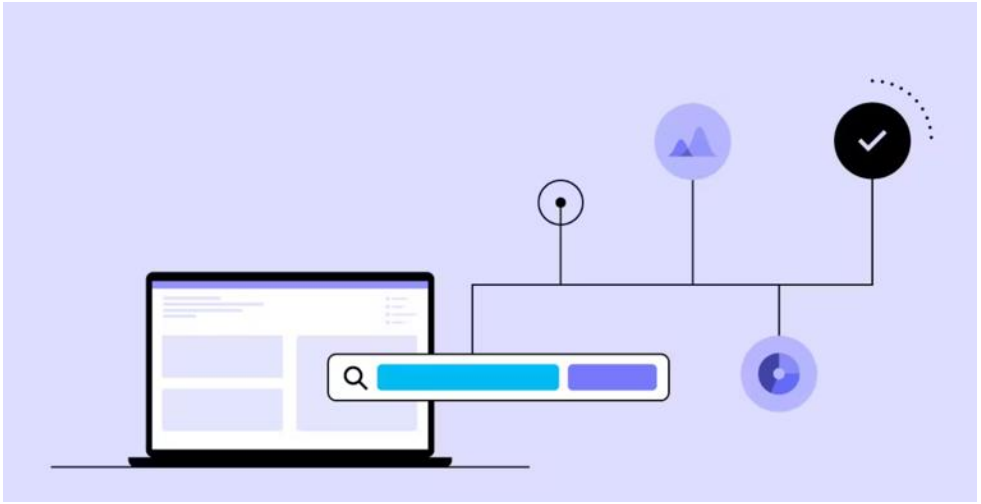
Data visualization principles

- **Avoid clutter:**
  - If you try to put too much information in the same visualization, the actionable information gets lost in the noise.
  - Instead, it's better to encourage the viewer to discover information through progressive disclosure–revealing each piece of information as you go to build a complete picture.
  - For instance, start by just showing the highest level data in your presentation and progressively drill down into the data to answer questions from your audience
- **5 categories of clutter:**
  - **Frame:** The background area of the chart.
  - **Graphical objects**: The elements that are used to visualize or display the data values.
  - **Axes**: The reference lines along the edges of the frame.
  - **Labels**: All text elements that describe the chart's data values, axes, etc.
  - **Overlay objects**: All elements that are overlaid or inserted into the chart for explanatory purposes.

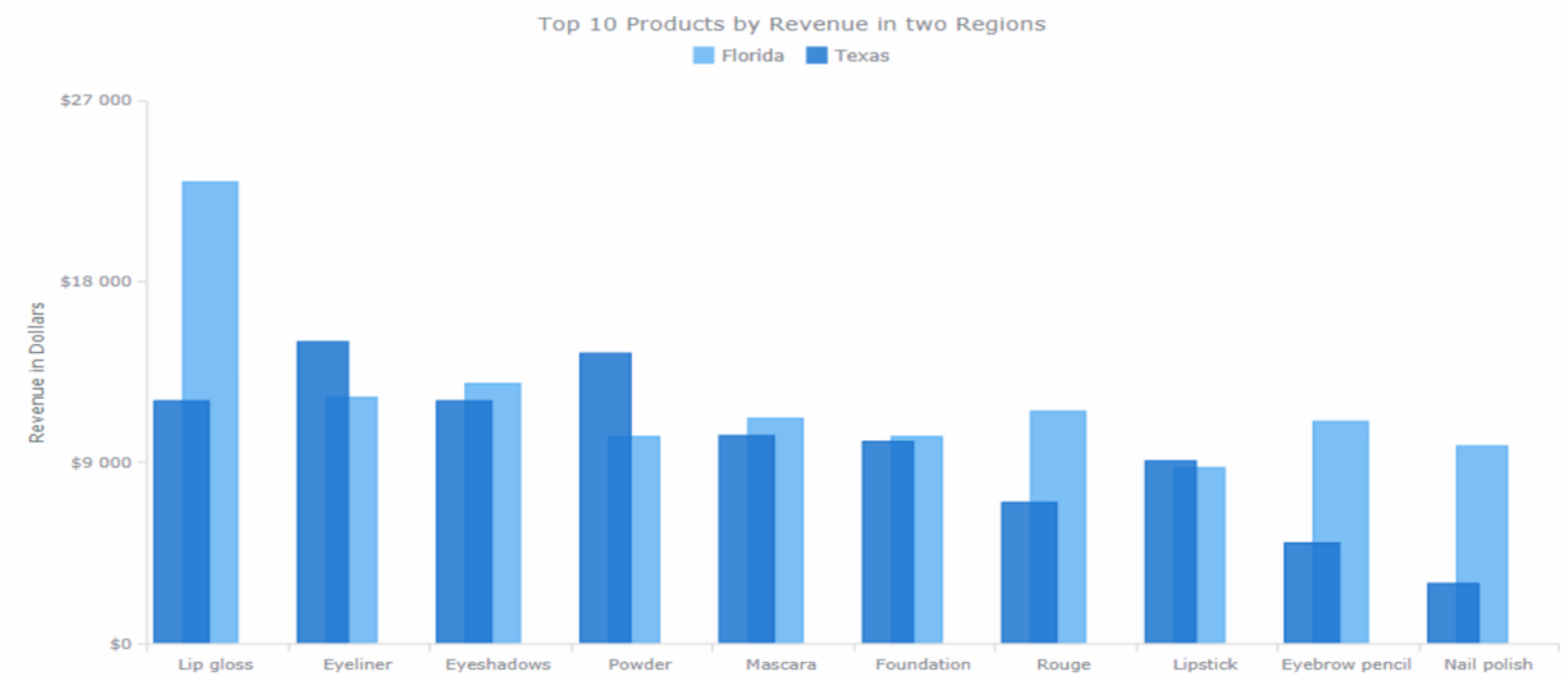


Data Visualization Approaches & Examples

- With dozens of chart types available, each known to suit one particular purpose or another, it is extremely important to choose the right chart type. In fact, picking a wrong form is one of the most common — and critical — visualization mistakes.
- Let's look at seven common examples of data visualization tasks and see what chart types are the best fits.
  - Compare Data
  - Explore composition & part-to-whole relationship in data
  - Track data over time
  - Analyze data distribution
  - Evaluate current performance of data
  - Examine project data
  - Make sense of geographical data



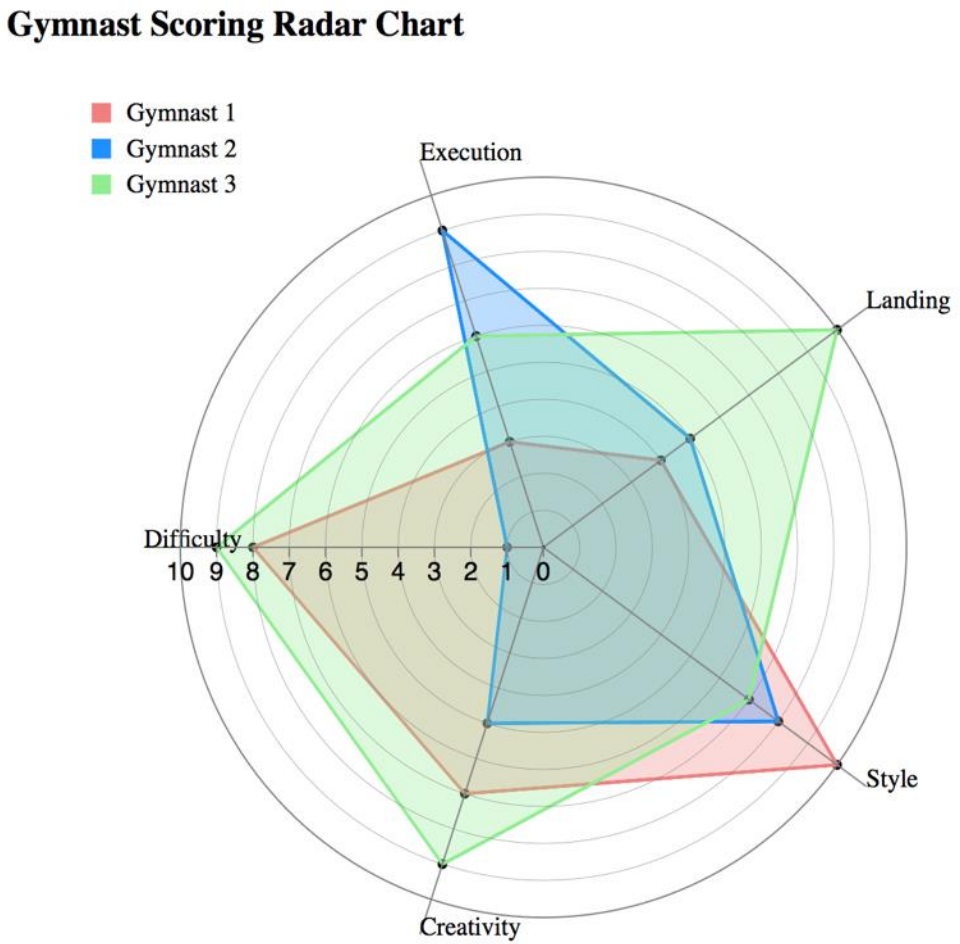
Compare data



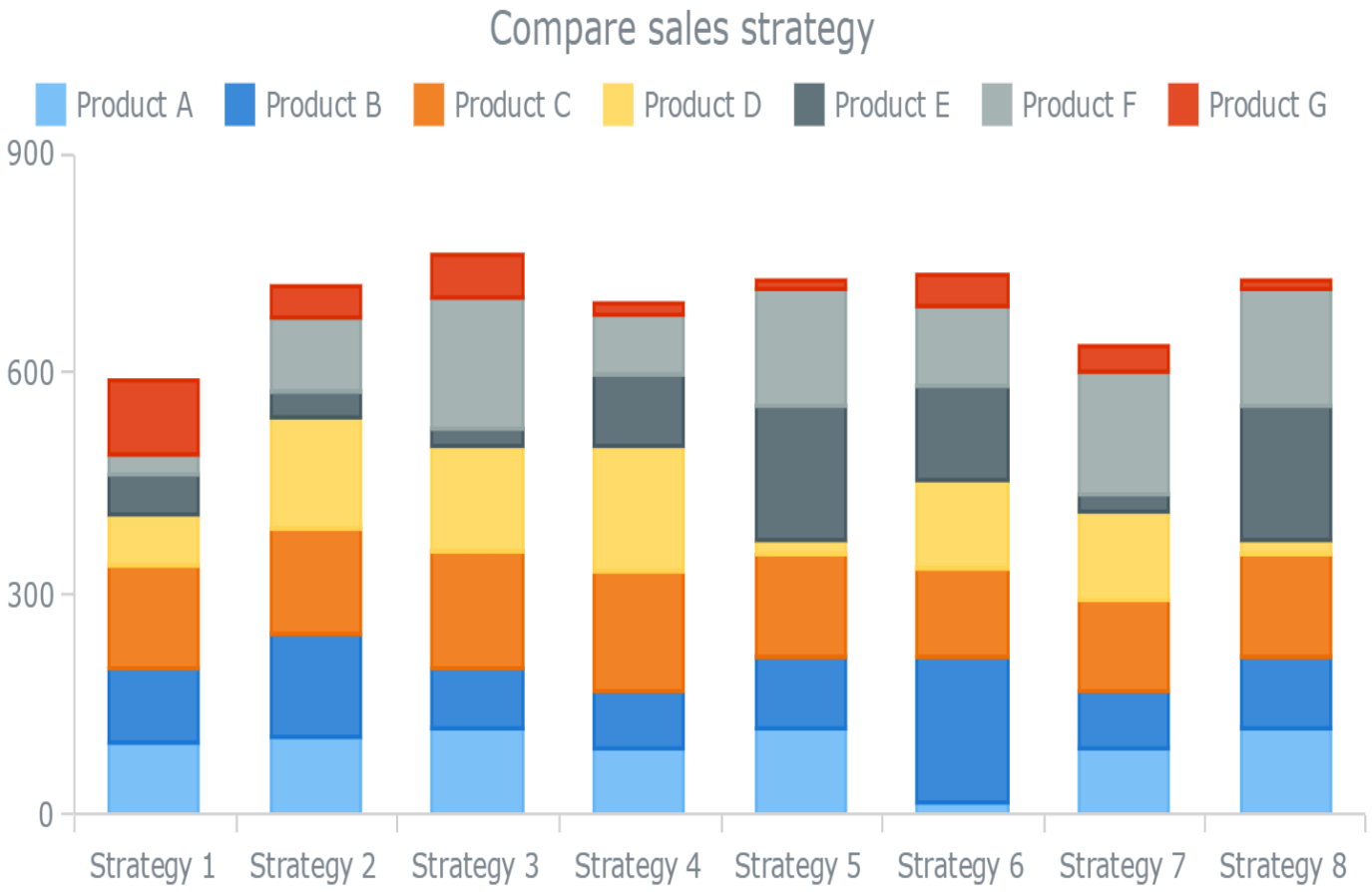
To compare data, the following chart types are commonly helpful:

- **Bar charts (and column charts)** — for a straightforward comparison of quantitative values by category.
- **Stacked charts** (and 100% stacked charts) — to add a look at the composition.
- **Radar charts** — for a comparison of cyclic data.
- For example, you can use bar charts to visualize products by revenue and referral sources by user traffic; stacked charts — sales figures by product and by region; radar charts — air temperature by month; etc.

Compare data

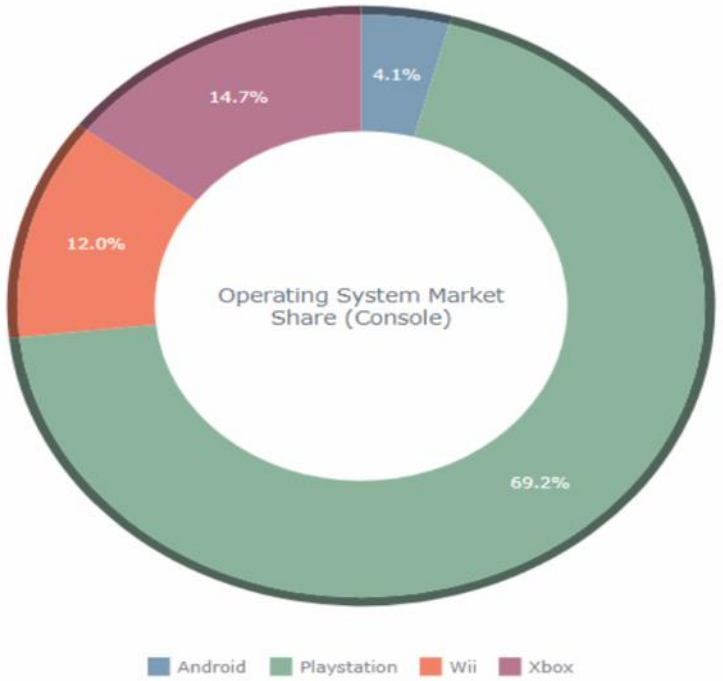


Radar charts



Stacked charts

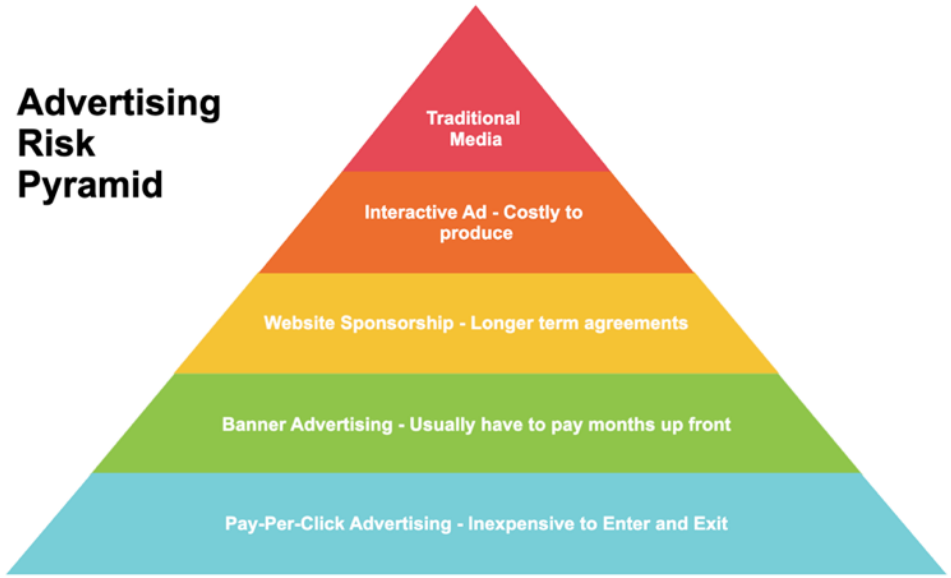
Explore Composition and Part-to-Whole Relationships in Data



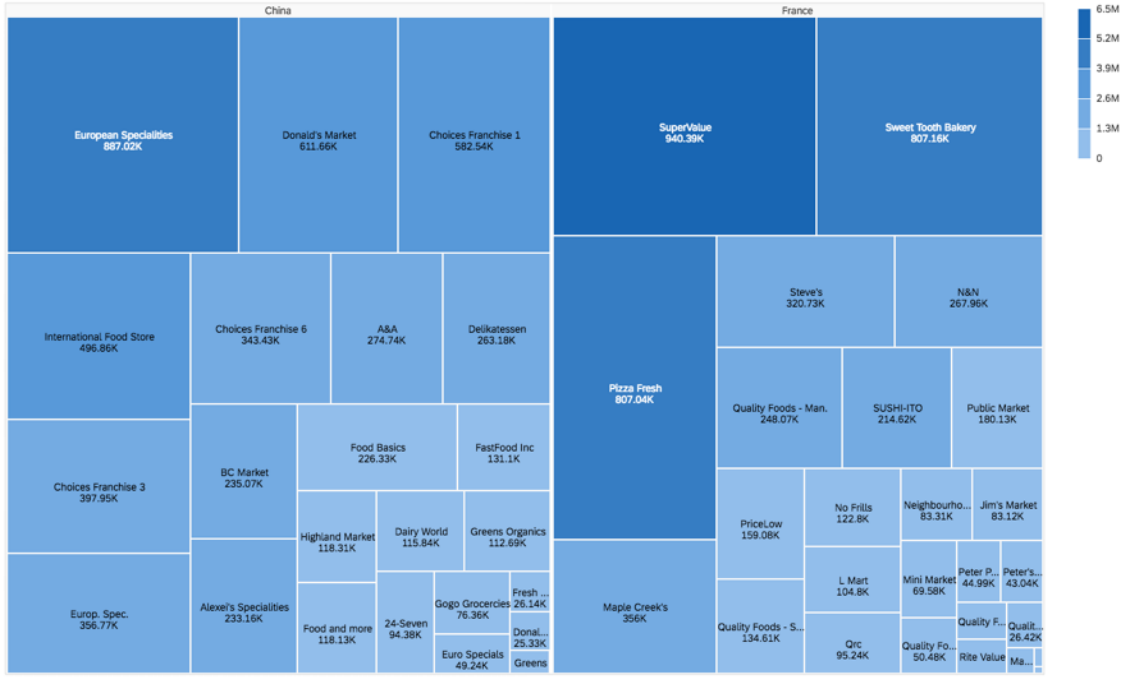
To explore the composition and part-to-whole relationships, the following chart types are commonly helpful:

- **Pie charts (and donut charts)** — for a basic look at the percentage composition of a value.
- **Pyramid charts** — to explore the composition of hierarchical data.
- **Treemap charts** — to look into complex hierarchical data.
- **Funnel charts** — to measure stages of processes and discover bottlenecks.
- pie or donut charts to visualize a breakdown of the overall sales by retail channel or one of the total website traffic by user age; pyramid charts — OSI model or employee salary by management level; treemap charts — web traffic by source category (and by flow amount) or export directions by country of destination; funnel charts — sales funnels; etc.

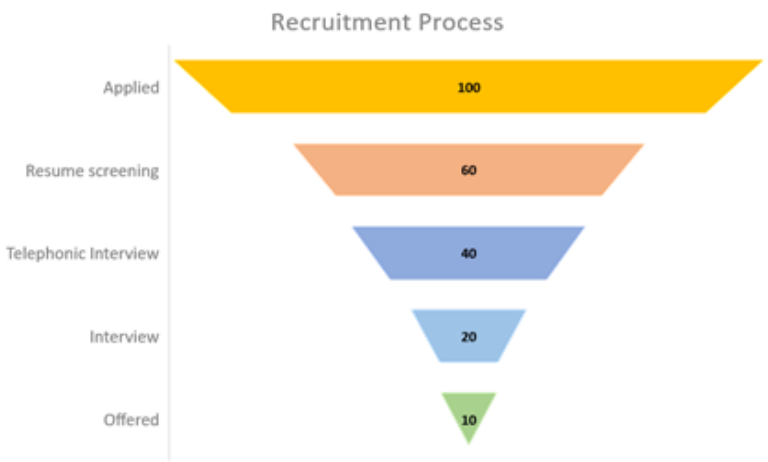
Explore Composition and Part-to-Whole Relationships in Data



Pyramid chart

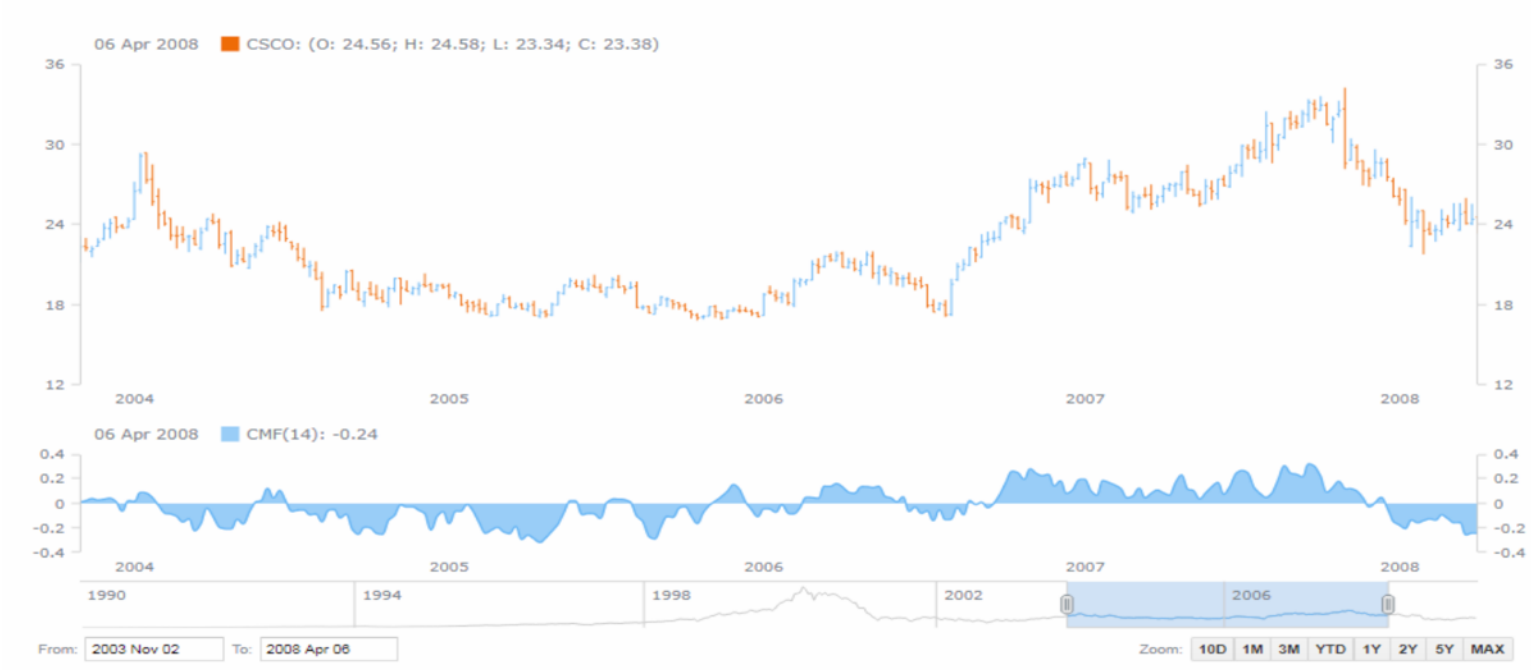


Tree map chart



Funnel chart

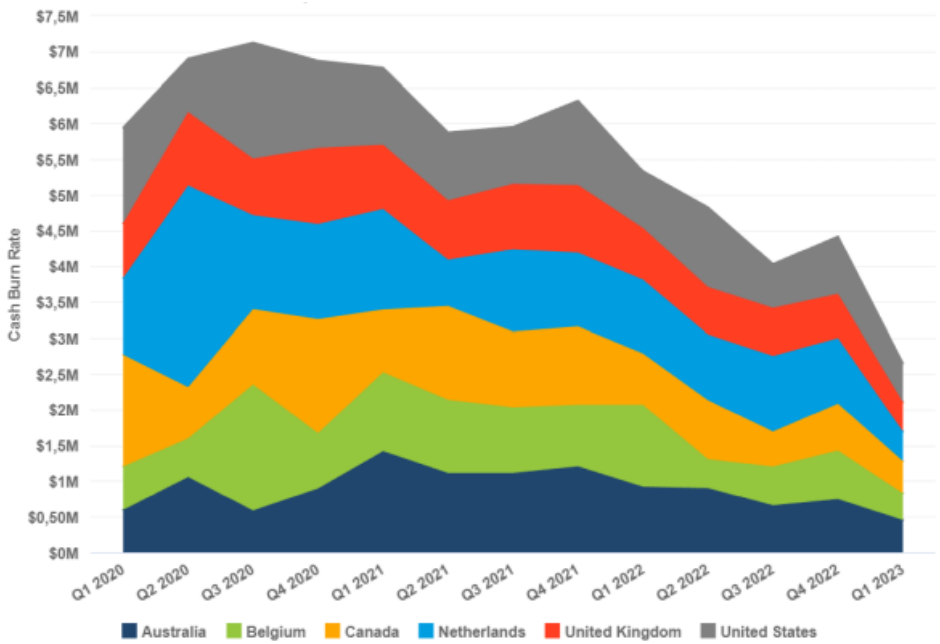
Track Data Over Time

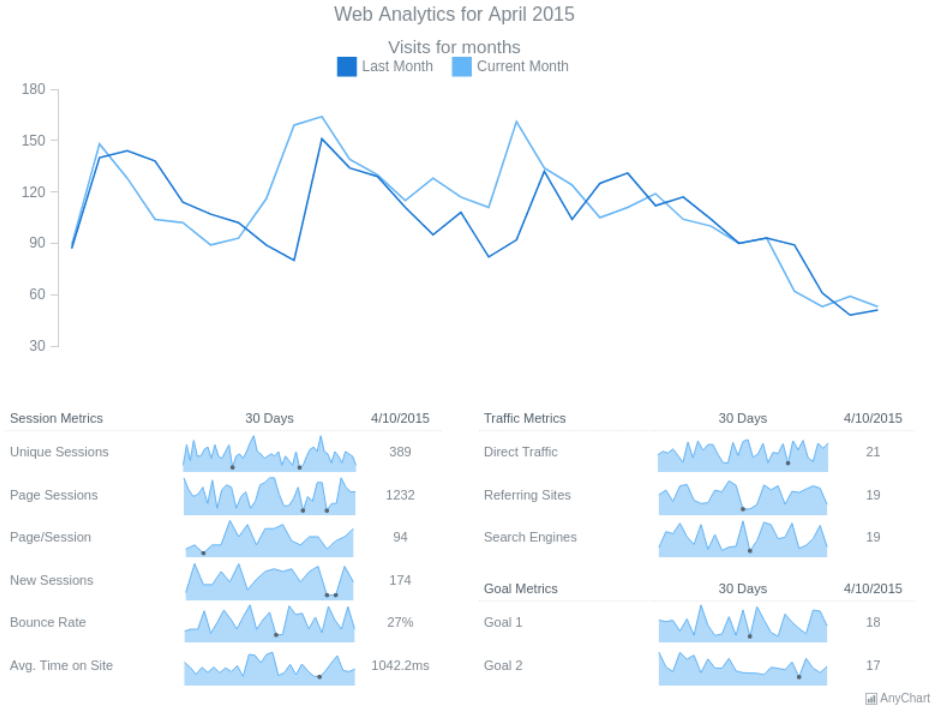


To track data over time, the following chart types are commonly helpful:

- **Line charts (and spline charts)** — for a basic view revealing trends, peaks, and so on.
- **Area charts** — as another option, e.g. for cumulative data.
- **Stock charts** — for big data sets such as financial and stock market data.
- **Candlestick charts (and OHLC charts)** — to add a look at the distribution of values within each period of time.
- **Sparkline charts** — for a quick representation of the big picture, with no axes.
- line, spline or area charts to visualize sales or web traffic over time; stock charts — stock price change (also along with technical indicators); OHLC and candlestick charts — stock price with a look into value fluctuation ranges within each of the numerous time periods; sparkline charts — overview of the sales performance within the last 12 months or football season win/loss results; etc.

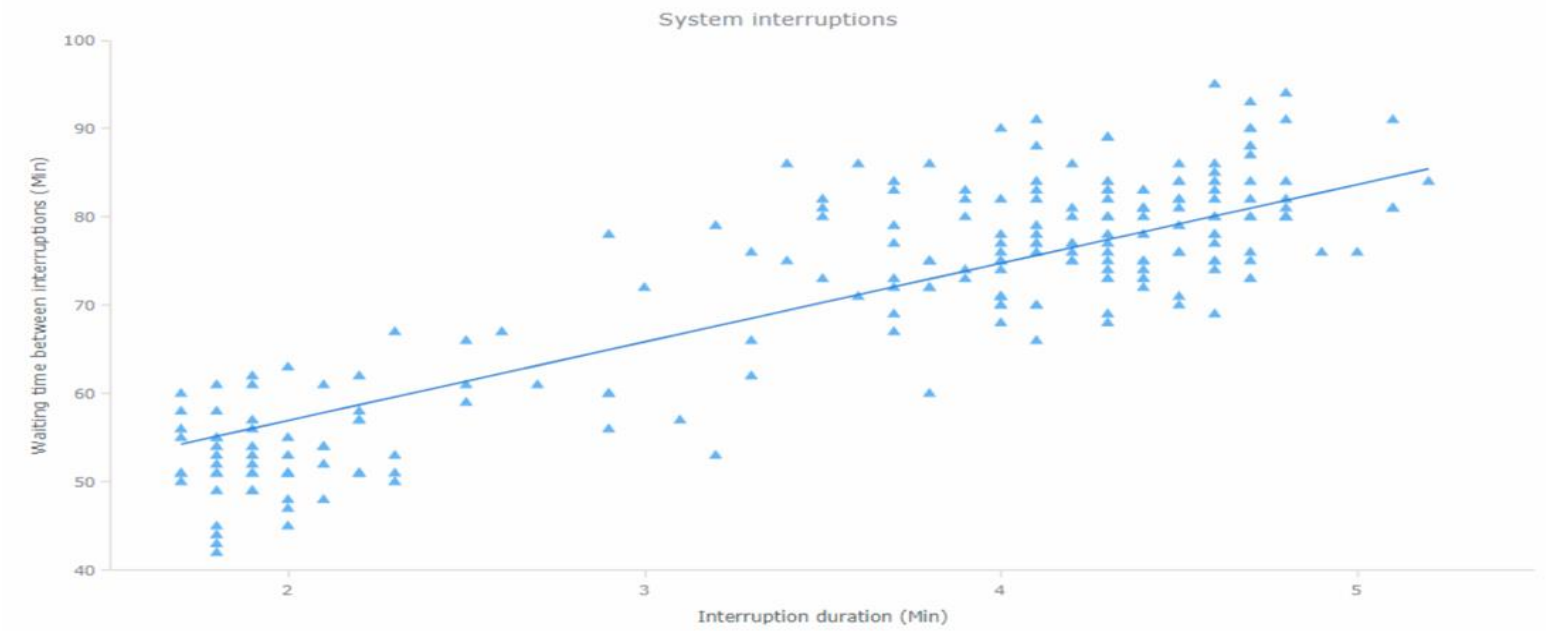
Track Data Over Time





Area, sparkline and candlestick charts

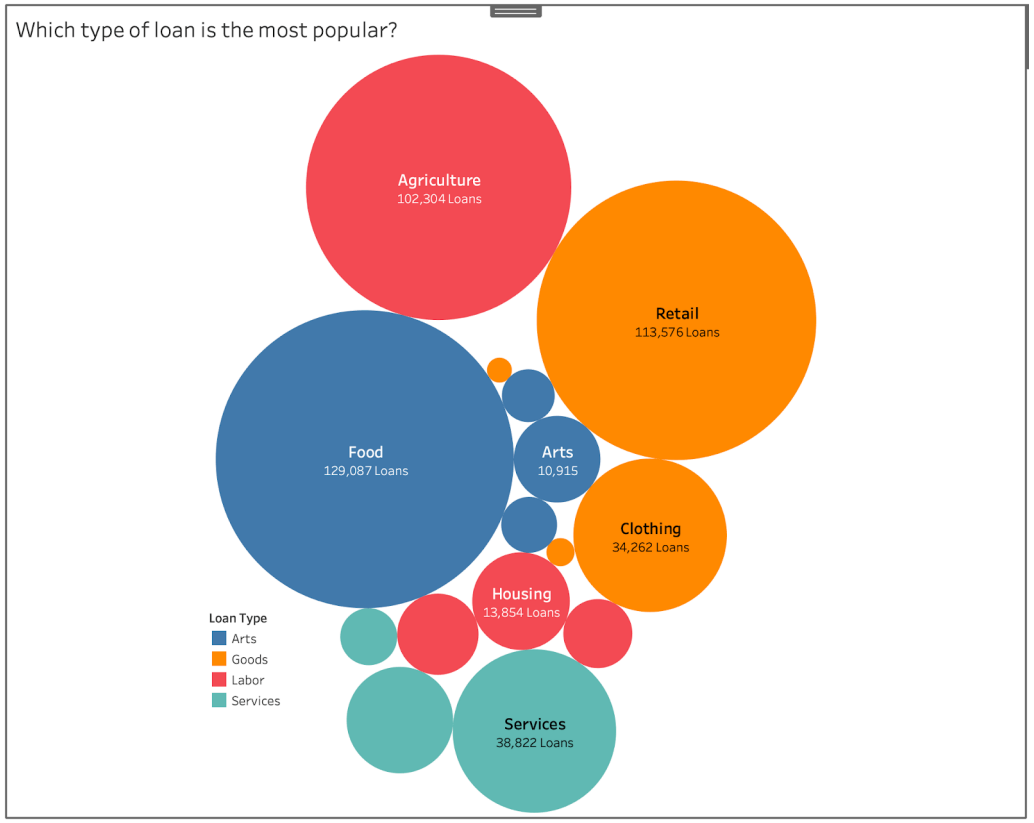
Analyze Data Distribution

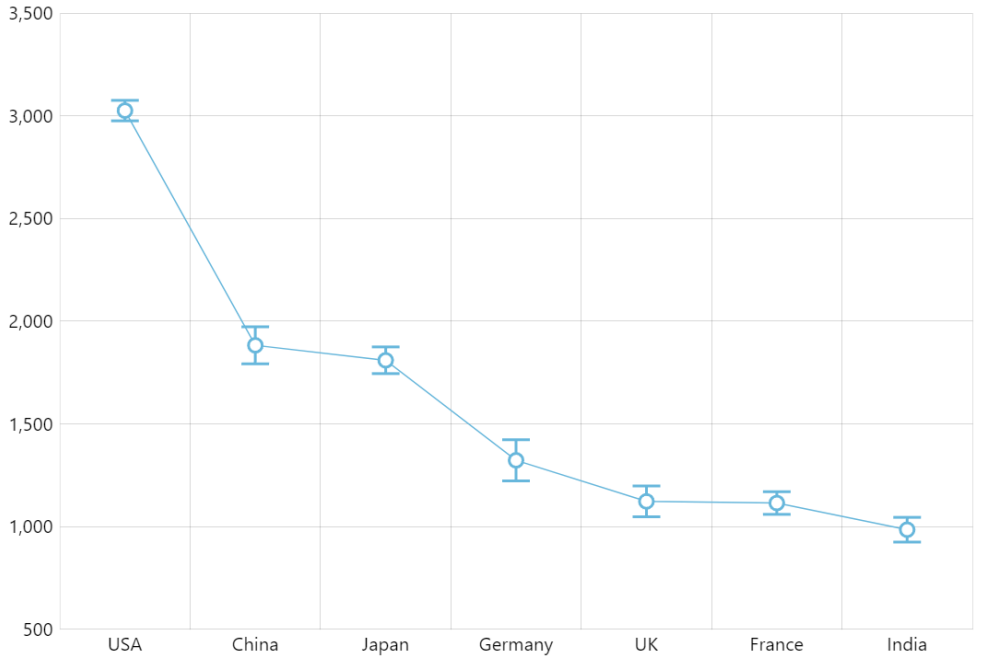


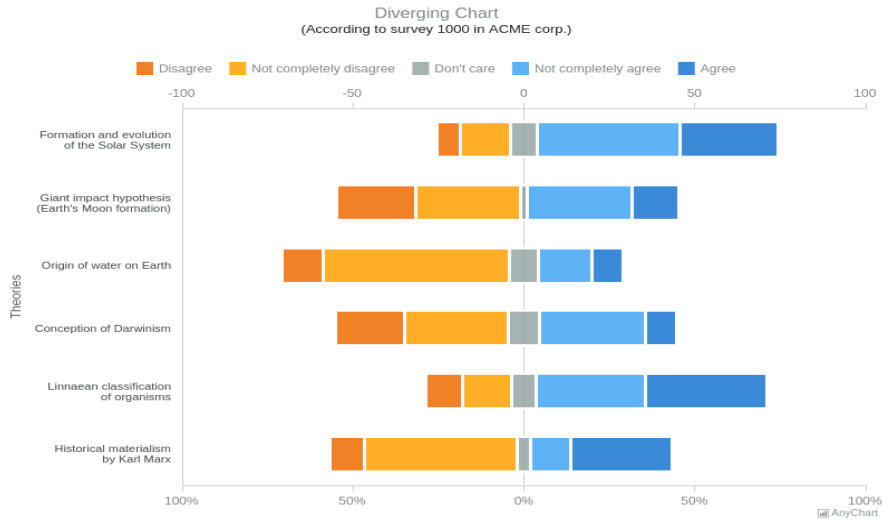
To analyze data distribution and relationship between data sets, the following chart types are commonly helpful:

- **Dot (scatter) charts** — to examine trends in distribution and correlation between two variables.
- **Bubble charts** — to consider three dimensions of data.
- **Box-and-whisker charts** — for a look at main distribution ranges and median values.
- **Error charts** — to inspect error distribution.
- **Heat map charts** — for a colored matrix-based view of multiple subcategories.
- **Range charts** — to mind range between the maximum and minimum values.
- **Polar charts** — for multivariate data with a spatial perspective.
- dot/scatter charts to visualize system interruptions by waiting time and by duration, or results of an experiment; bubble charts — training data by sportsman, power, and pulse; box-and-whisker charts — destinations by flight delay duration; error charts — variability of product sales

Analyze Data Distribution

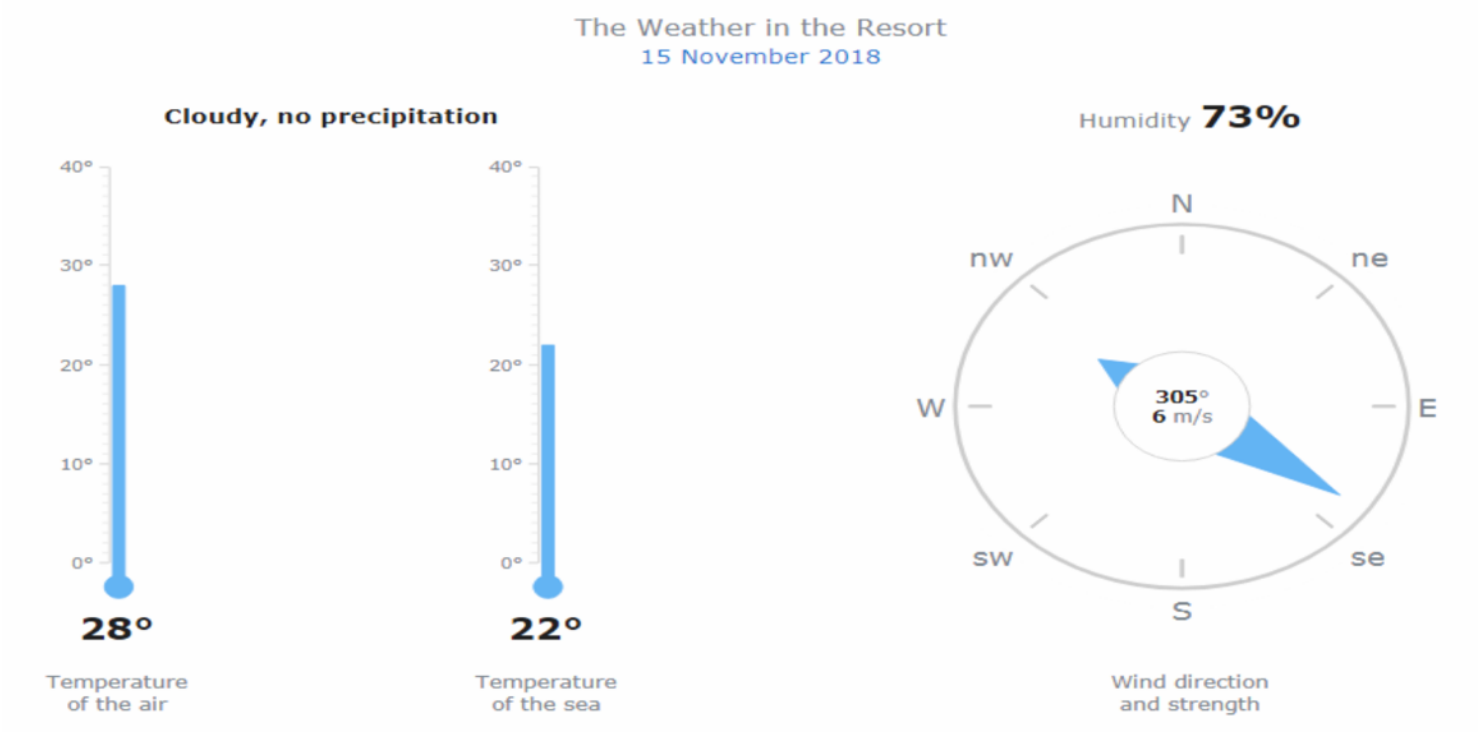






Bubble , error and range charts

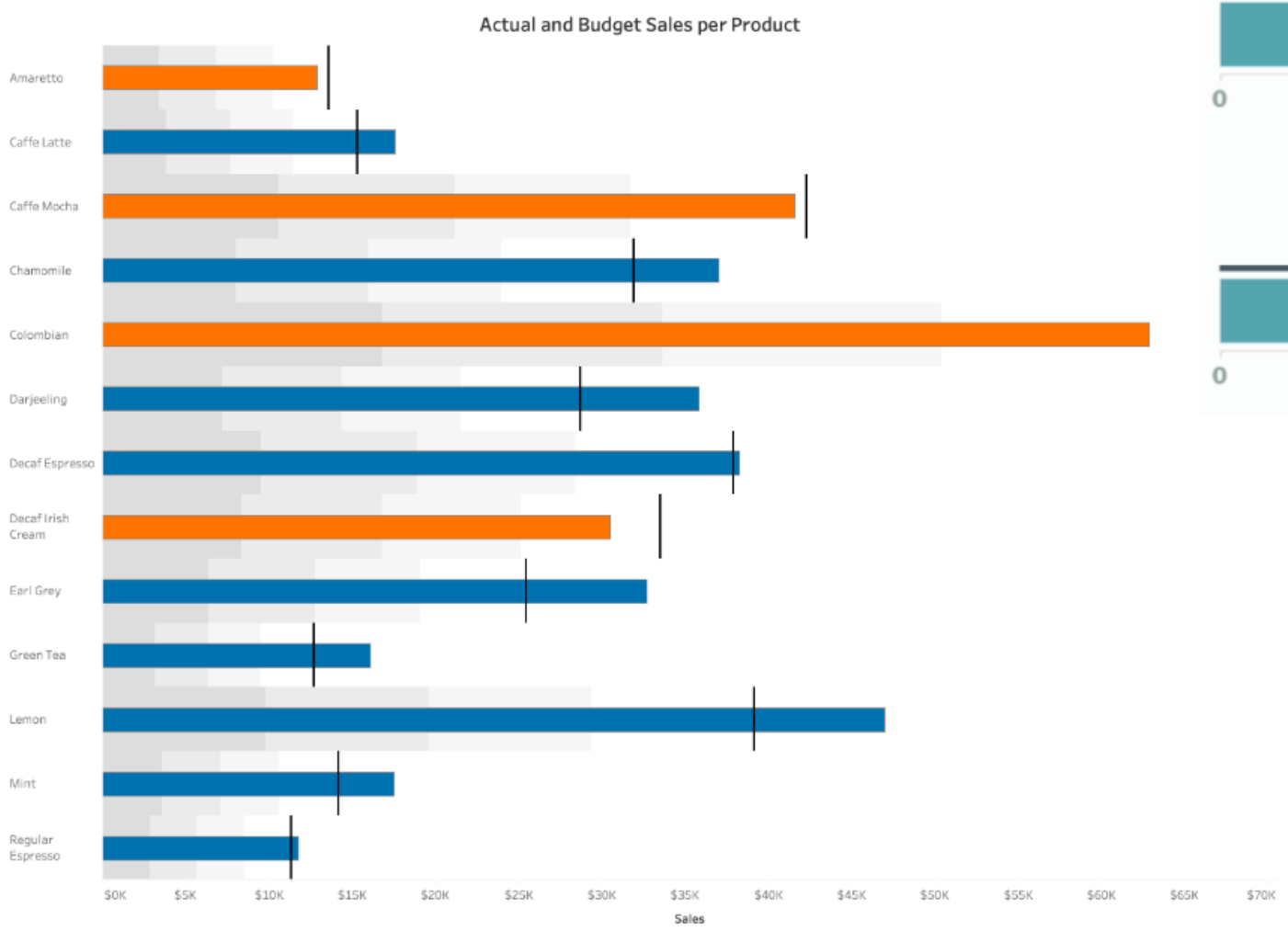
Evaluate Current Performance Data

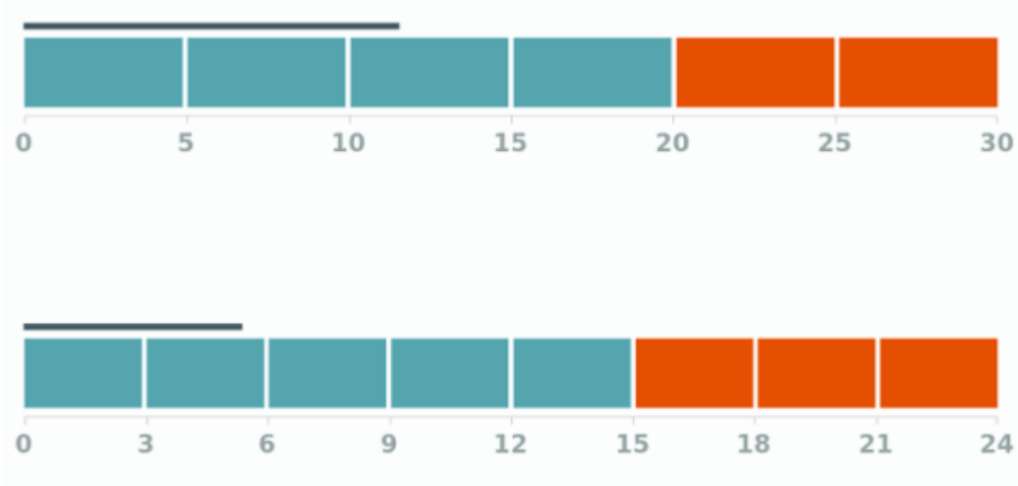


To evaluate performance and compare the actual value against KPI values or qualitative ranges — and basically, to visualize single-value data as indicators — the following chart types are commonly helpful:

- **Circular gauges** — for a radial scale.
- **Linear gauges** — for a linear scale.
- **Bullet charts** — for a more space-efficient visualization on a linear scale.
- For example, you can use circular gauges to visualize speed (like a speedometer) or wind direction (employing a compass-based look); linear gauges — air temperature (thermometer) or tank volume; bullet charts — current vs. planned sales or system availability figures; etc.

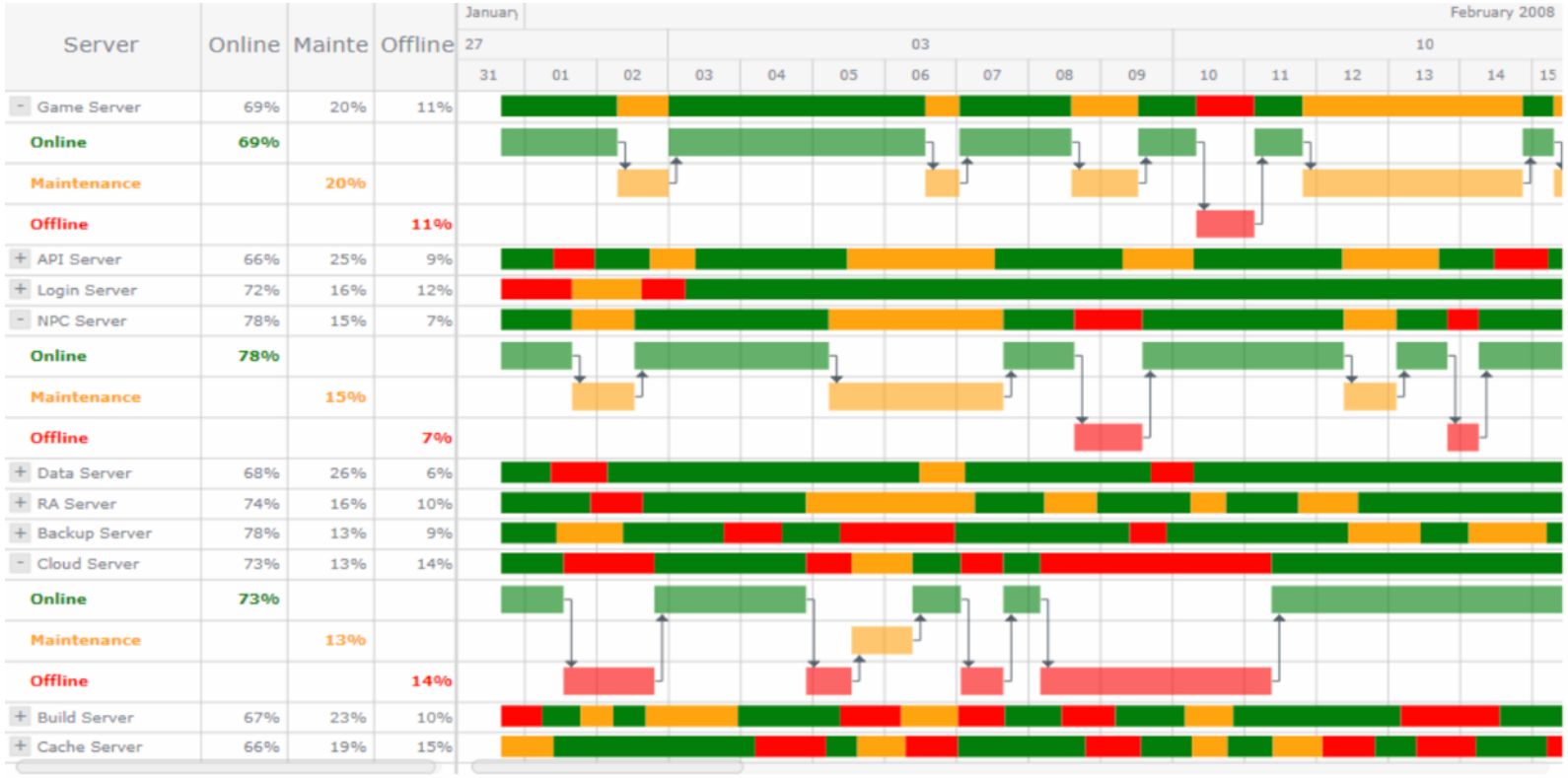
Evaluate Current Performance Data





Bullet charts, linear gauges, circular gauges

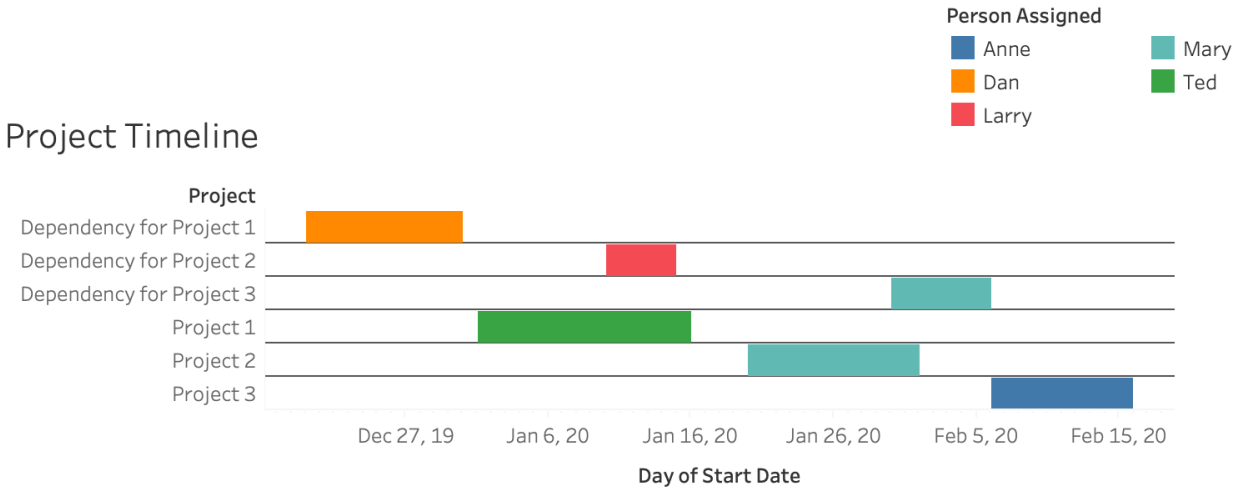
Examine Project Data

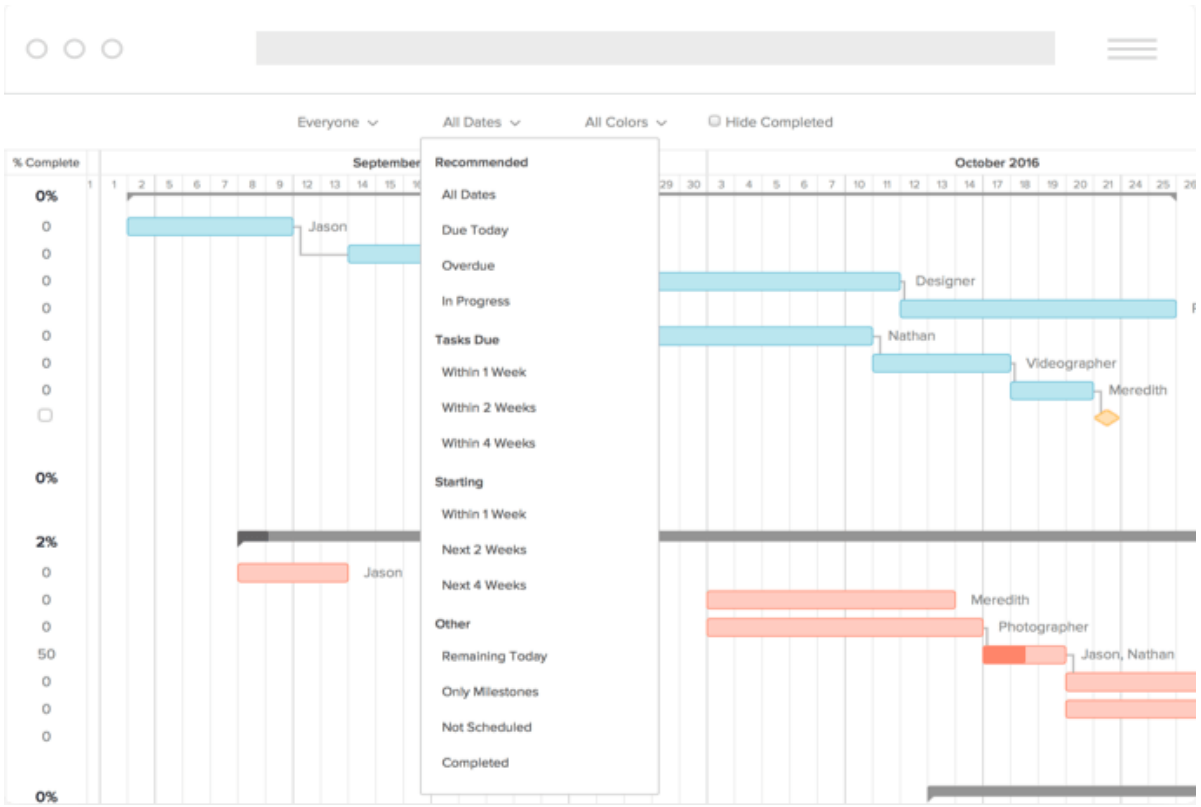


To examine project data, the following chart types are commonly helpful:

- **Gantt chart** — to keep an eye on activities on a project schedule.
- **Resource chart** — to review resource occupancy.
- For example, you can use Gantt charts to visualize project activities on a schedule or planned vs. actual progress; resource charts — server status; etc

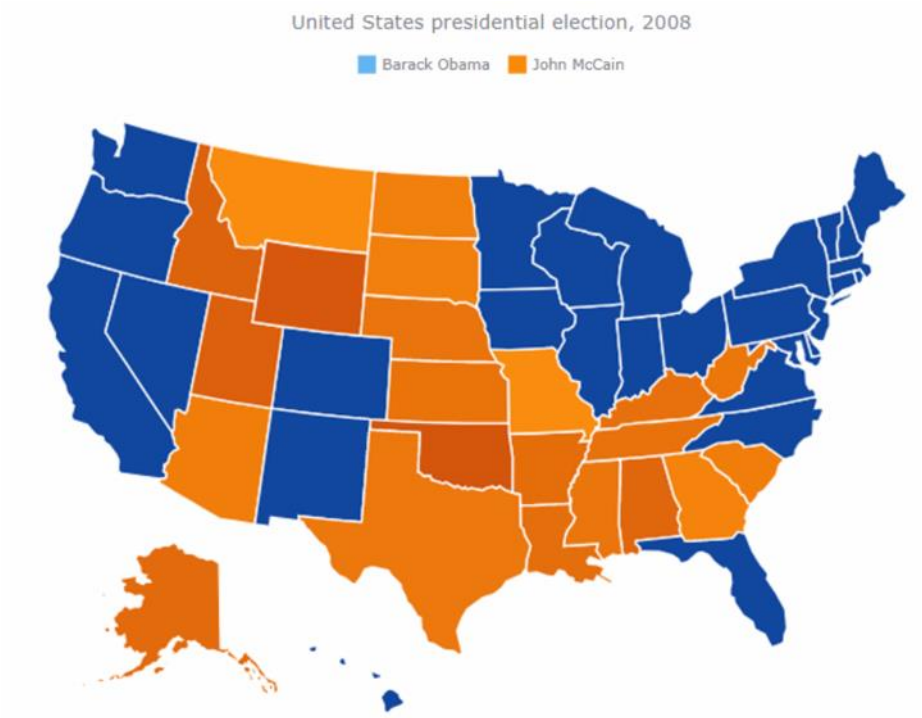
Examine Project Data





Gantt chart, resource chart

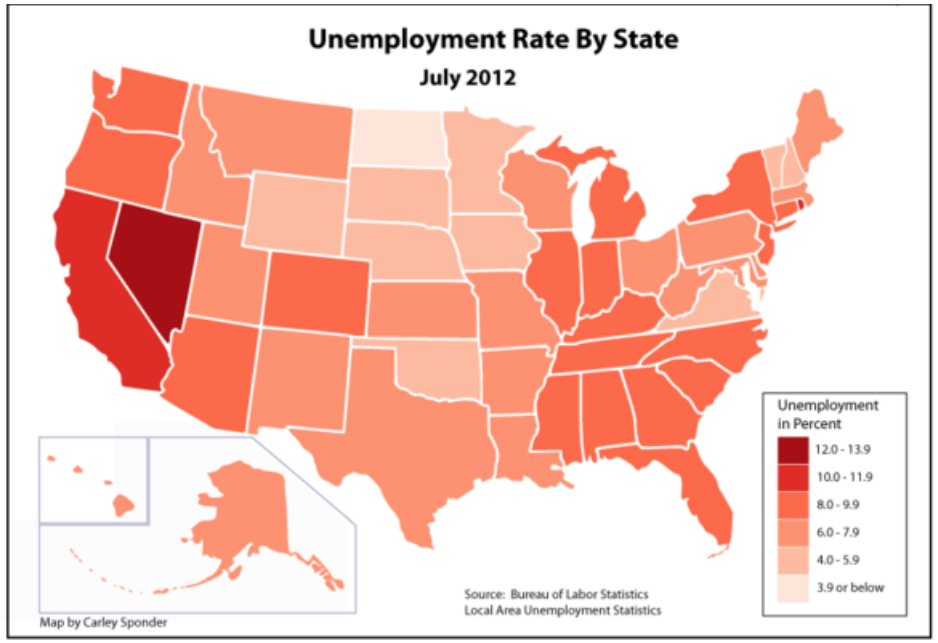
Make Sense of Geographical Data

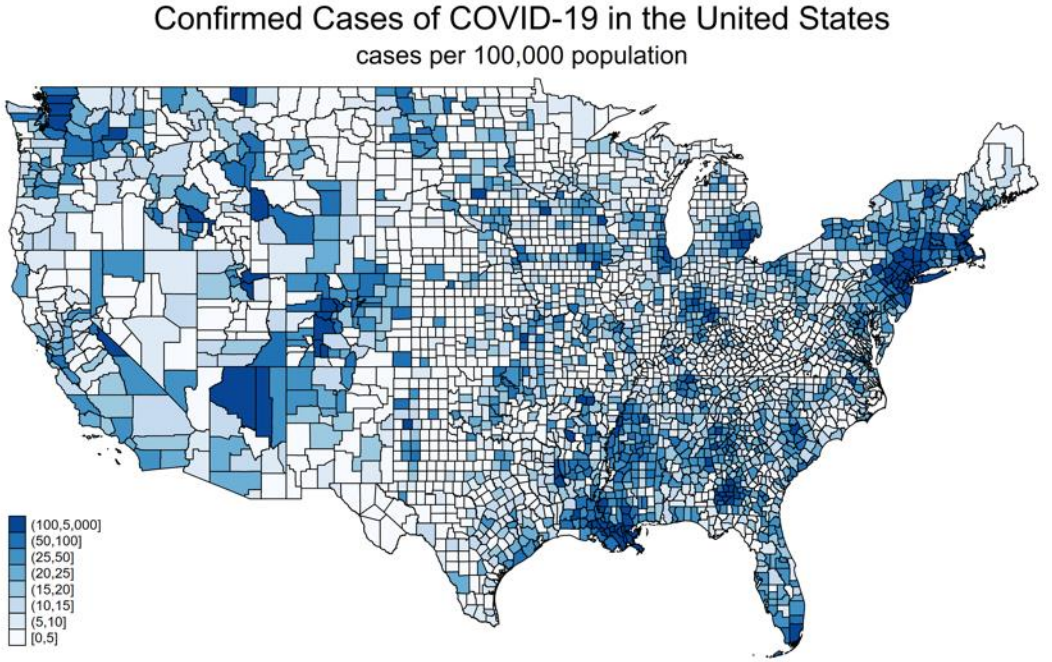


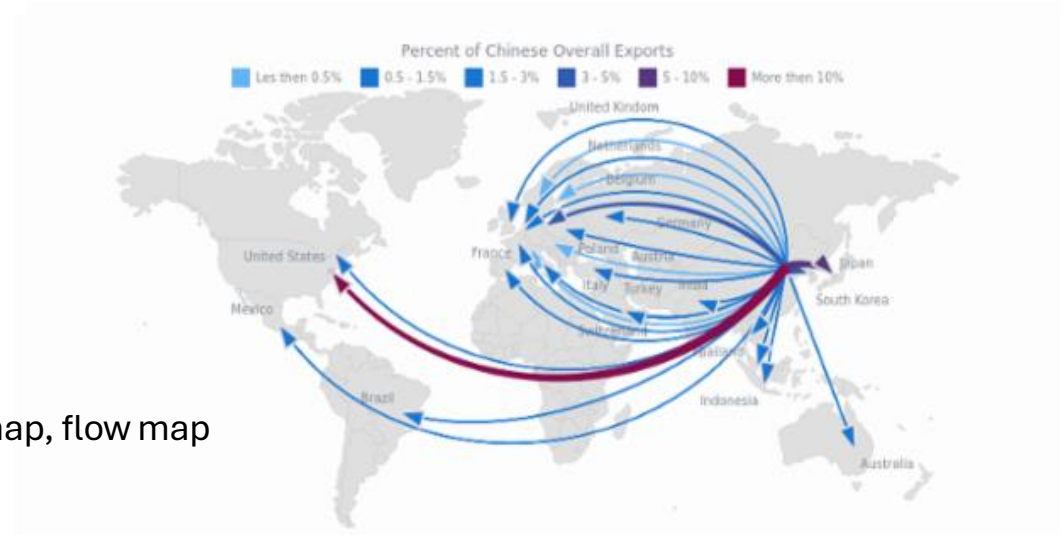
To make sense of geographical data, the following geo-visualization chart types are commonly helpful:

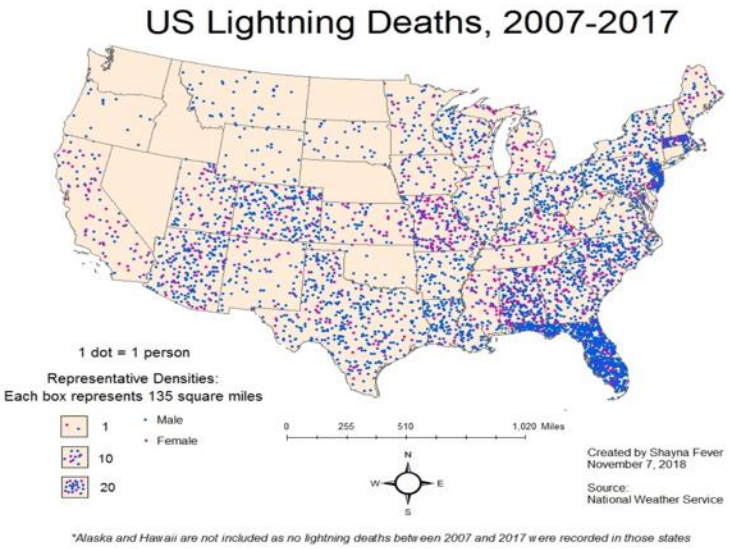
- **Choropleth maps** — to identify differences across geographical areas.
- **Dot maps** — to understand geographical distribution trends.
- **Bubble maps** — to add a size variable into a visual.
- **Connector maps** — for a look at geographical connections.
- **Flow maps** — to explore how objects move between locations when direction is important.

Make Sense of Geographical Data









Choropleth map, dot map, flow map

Different types of charts & graphs

- Line charts
  - A line chart connects distinct data points through straight lines. It’s used to find trends, patterns, and variable changes.

When to use line charts?

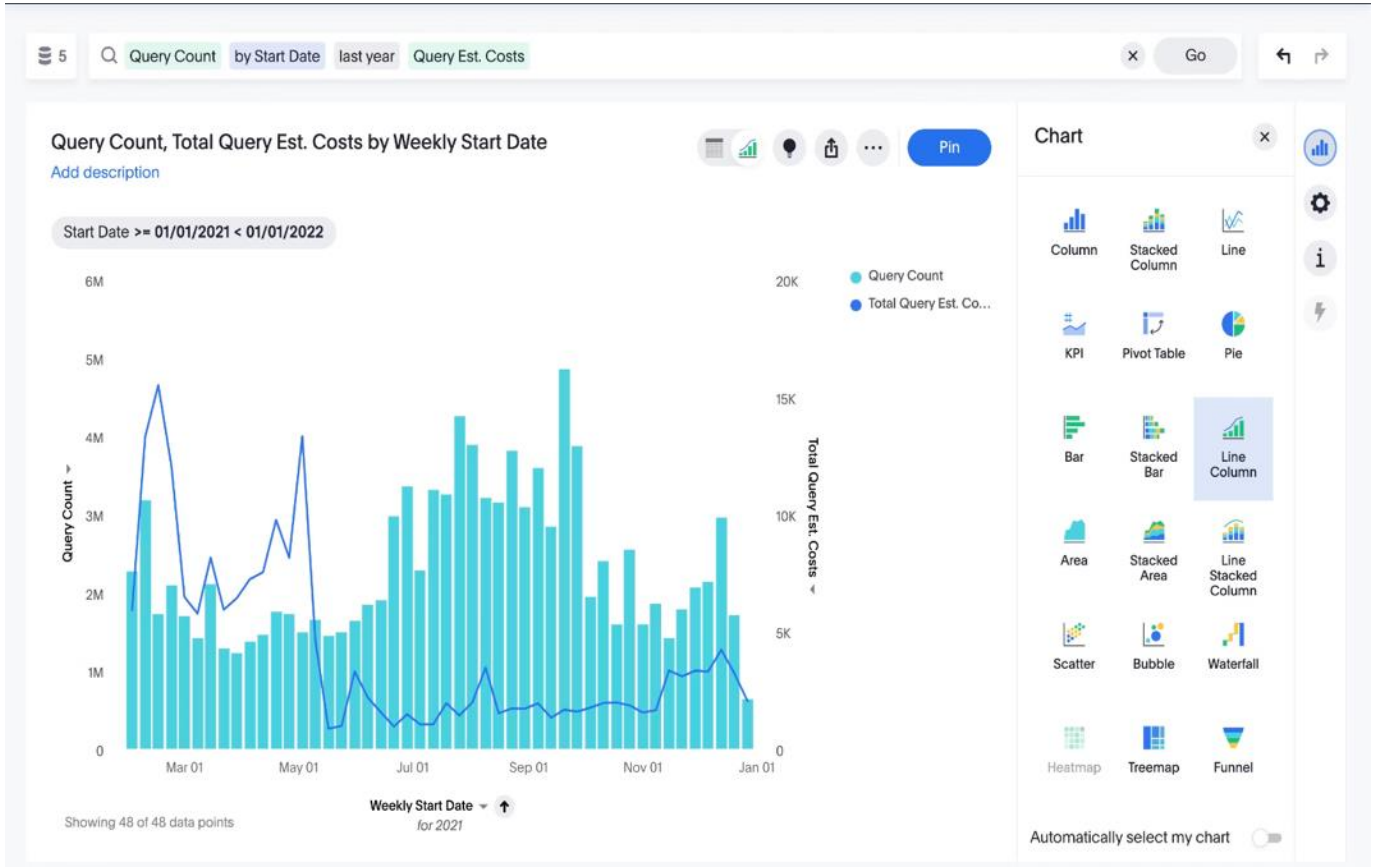
This type of chart helps measure how different groups relate to each other. This type of chart is also effective for demonstrating progression, making them suitable for scenarios like project timelines, production cycles, or population growth.

Best practices for line charts:

•Ensure that the data you're representing has a logical order

•Add context through annotations and labels

•If the dataset is large, use transparency or spacing to improve visibility



Bar Charts

- A bar chart represents data using rectangular bars or columns, the length of each bar corresponds proportionally to its value. You can present these bars horizontally or vertically.

When to use bar charts?

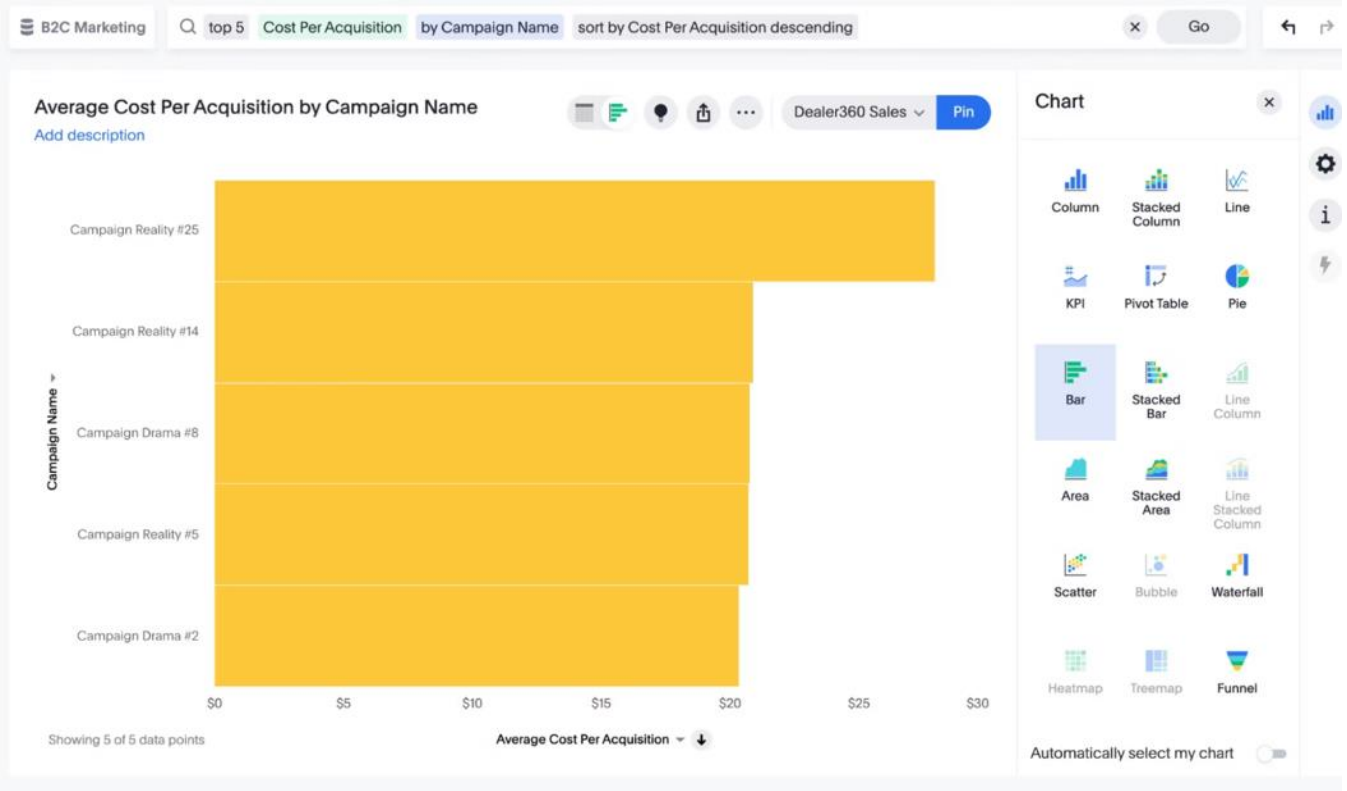
Bar charts are excellent for comparing the values of different categories or groups. They are also helpful in showing distribution of data across different categories.

Best practices for bar charts:

•Clearly label each bar and axis with concise labels

•Limit the number of bars and categories to avoid cognitive overload

•Purposely use colors to highlight key points and convey meaning



Treemap Charts

- Treemaps are hierarchical charts that allow to visualize data as nested rectangles. These rectangles or branches convey the structure and distribution of data, making treemaps useful for visualizing categorical and hierarchical relationships.

When to use treemap charts?

Apart from visualizing hierarchical data, this type of visualization helps to illustrate part-to-whole relationships within a dataset, demonstrating how each category contributes to the overall composition.

Best practices for treemap charts:

•Ensure the size of the rectangle is proportional to the size of the category

•Clearly label each rectangle with concise labels

•Use a single color with varying shades to show changes in data

Scatter Plots

- Scatter plots are types of visualization that show a collection of data points ‘scattered’ around the graph. The data points can be evenly or unevenly distributed.

When to use scatter plots?

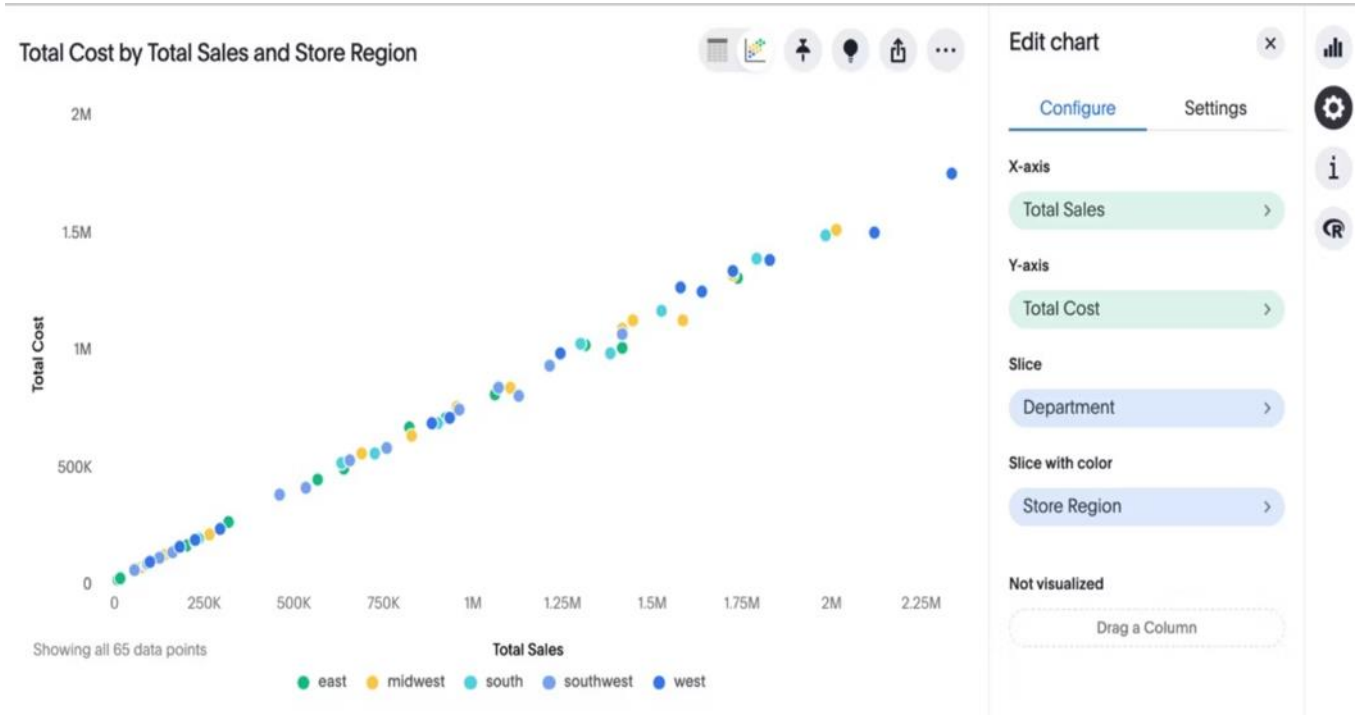
Scatter plots are ideal for exploring relationships and patterns between two continuous variables. They can help you identify trends, correlations, or potential clusters in the data.

Best practices for scatter plots:

•Highlight outliers if present in the graph to showcase data distribution

•Add a trendline to highlight the relationship between variables

•Consider using different colors or marker sizes for overlapping points



Heatmap Charts

- Heatmap charts are a type of map data visualization that uses a system of color coding to represent value.

When to use heatmap charts?

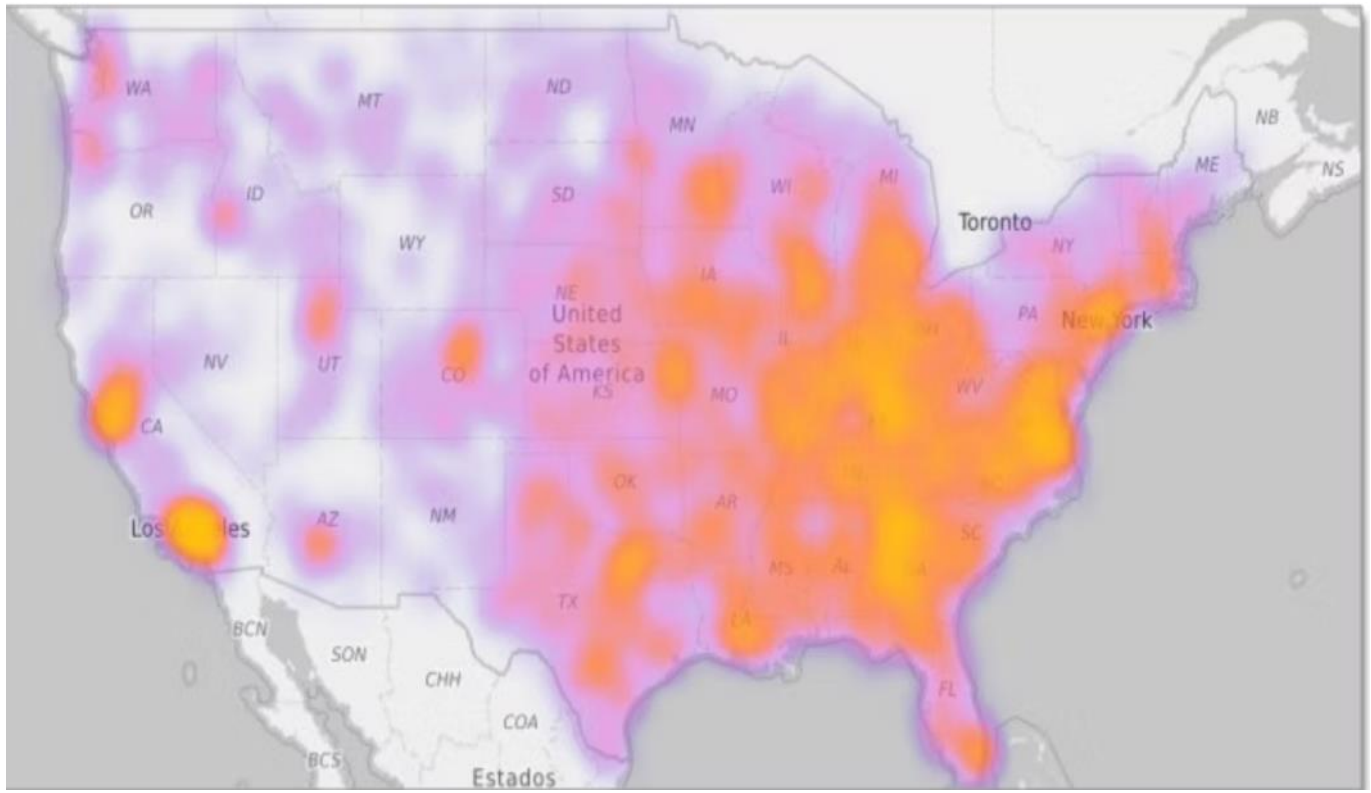
A heatmap is commonly used to establish relationships between two variables across a grid. The intensity of colors demonstrates the variables, making it easy to identify patterns and trends.

Best practices for heatmap charts:

•Choose an intuitive color palette that effectively conveys the magnitude of values

•Use visual cues to highlight significant values

•Utilize the design principle of white space to prevent overcrowding



Pareto Charts

- A Pareto chart combines a bar chart and a line graph. The rectangular bars correspond to individual values in descending order, while the line graph displays the cumulative percentage total. This type of chart follows the famous Pareto principle that emphasizes that ***20 percent of causes result in 80 percent of problems***

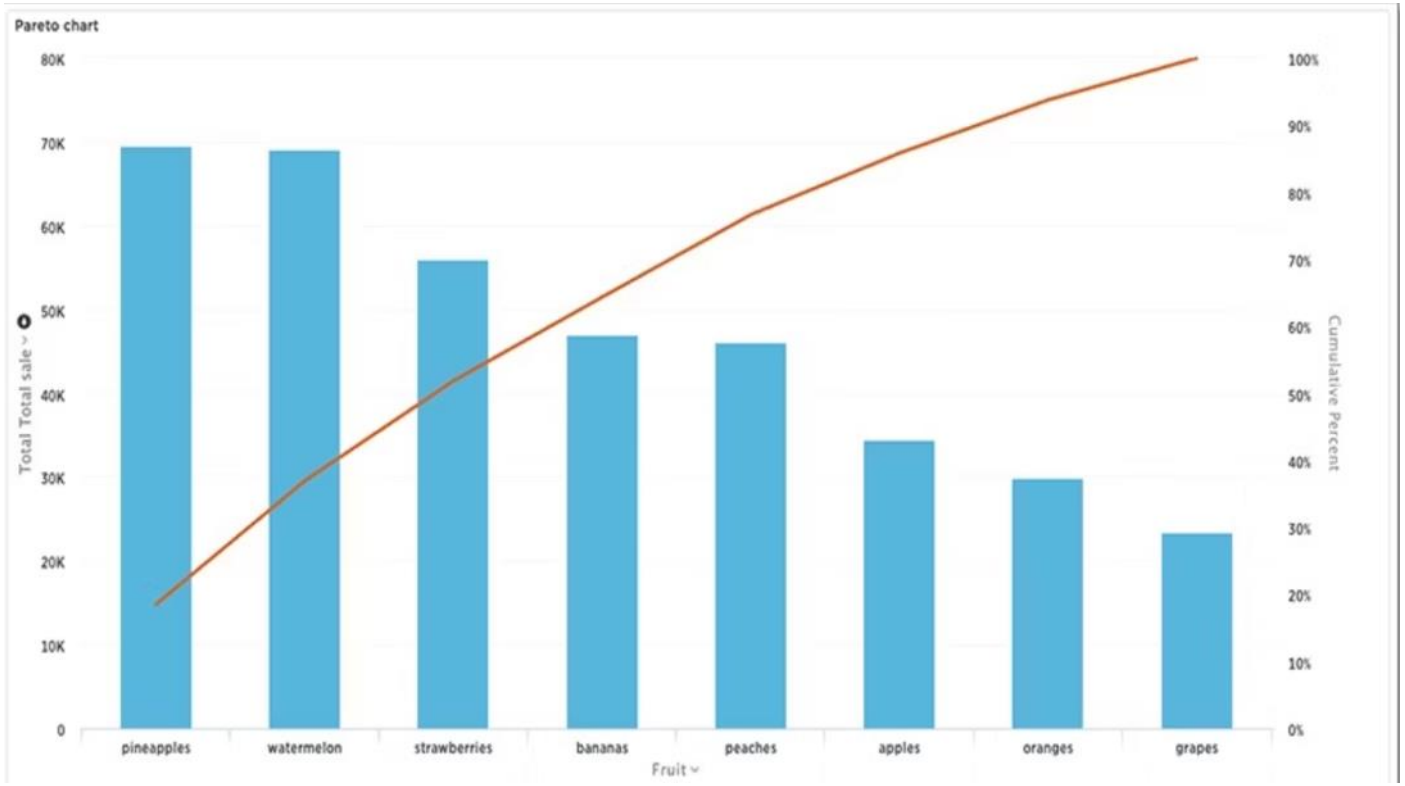
When to use Pareto charts?

A Pareto chart effectively showcases key contributing factors to a particular outcome. Another use case of a Pareto chart is when you want to highlight problems based on their impact.

Best practices for Pareto charts:

•Arrange all the categories in descending order based on their frequency, impact, or contribution

•Use colors purposefully to enhance clarity



Geo Charts

- Geo charts are a type of visualization that represent data on a map. They show spatial information, such as the distribution of values across different regions, countries, or states.

When to use geo charts?

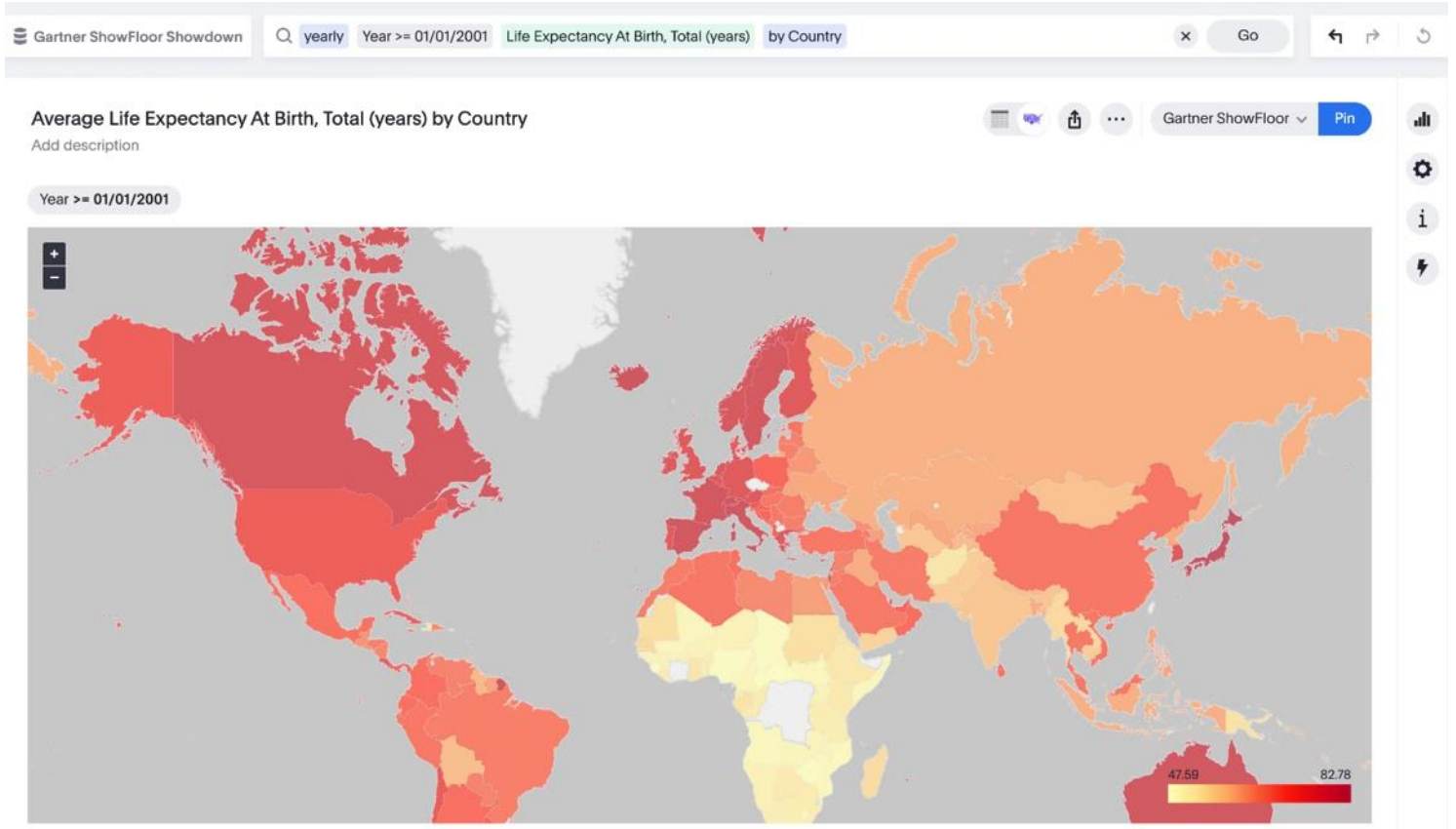
If you want to analyze geographic information in your data, you can use these types of charts to discover hidden patterns and trends. Each region, such as a country, state, or district, is shaded or colored based on the magnitude of the variable presented.

Best practices for geo charts:

•Select appropriate map projection that matches your region

•Use color shading to highlight particular regions

•Label the projections to represent data points on the map



What is Business Intelligence

- Business intelligence combines business analytics, data mining, data visualization, data tools and infrastructure, and best practices to help organizations make more data-driven decisions. In practice, you know you've got modern business intelligence when you have a comprehensive view of your organization's data and use that data to drive change, eliminate inefficiencies, and quickly adapt to market or supply changes.

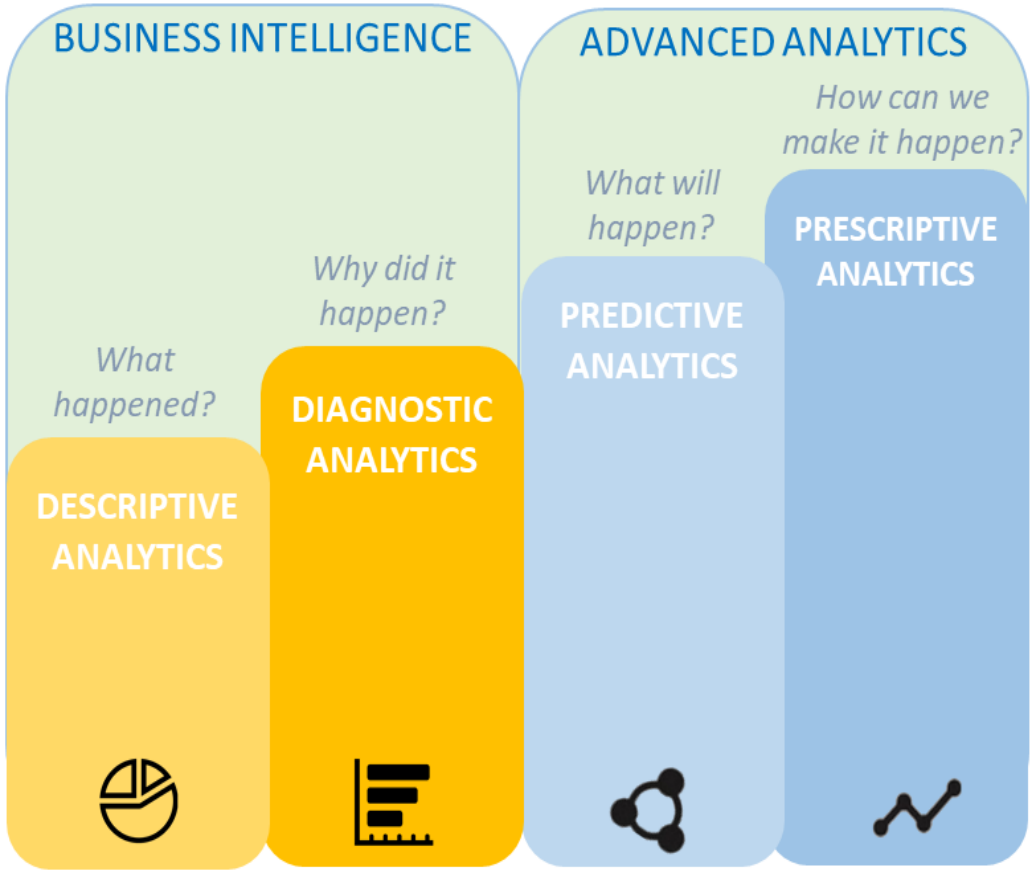
- Business intelligence is an infrastructure that helps in the process of collecting, storing, and analyzing data from business operations. BI focuses on descriptive analytics

- BI provides comprehensive business metrics, in near-real-time, to support better decision making. You can create performance benchmarks, spot market trends, increase compliance, and improve almost every aspect of your business with better business intelligence

https://www.tableau.com/learn/articles/business-intelligence

TYPES OF ANALYTICS

- **Descriptive analytics** – describes the past status of the domain of interest using a variety of tools through techniques such as reporting, data visualization, dashboards, and scorecards
- **Predictive analytics** – applies statistical and computational methods and models to data regarding past and current events to predict what might happen in the future
- **Prescriptive analytics** –uses results of predictive analytics along with optimization and simulation tools to recommend actions that will lead to a desired outcome



What is business analytics & data analytics

- A subset of BI, business analytics (BA) refers to a process of using raw data to find trends and predicting outcomes. Some common methodologies in business analytics are:
  - **Data mining:** sorting through large amounts of data to identify patterns and trends
  - **Aggregation:** the process of gathering and organizing data prior to analysis
  - **Forecasting:** analyzing historical data estimate future outcomes
  - **Predictive modeling:** extracting information from data sets to identify patterns and trends
  - **Data visualization:** creating visual representations of data analysis, such as charts, tables, or graphs

- BA focuses on predictive analytics
- Data analytics is a technical process of mining data, cleaning data, transforming data, and building the systems to manage data. Data analytics takes large quantities of data to find trends and solve problems. Data analytics is not just confined to business applications—it's used across disciplines, from government to science.

How Business Intelligence works

Data – raw material; has volume, velocity, and variety

Information – actionable data

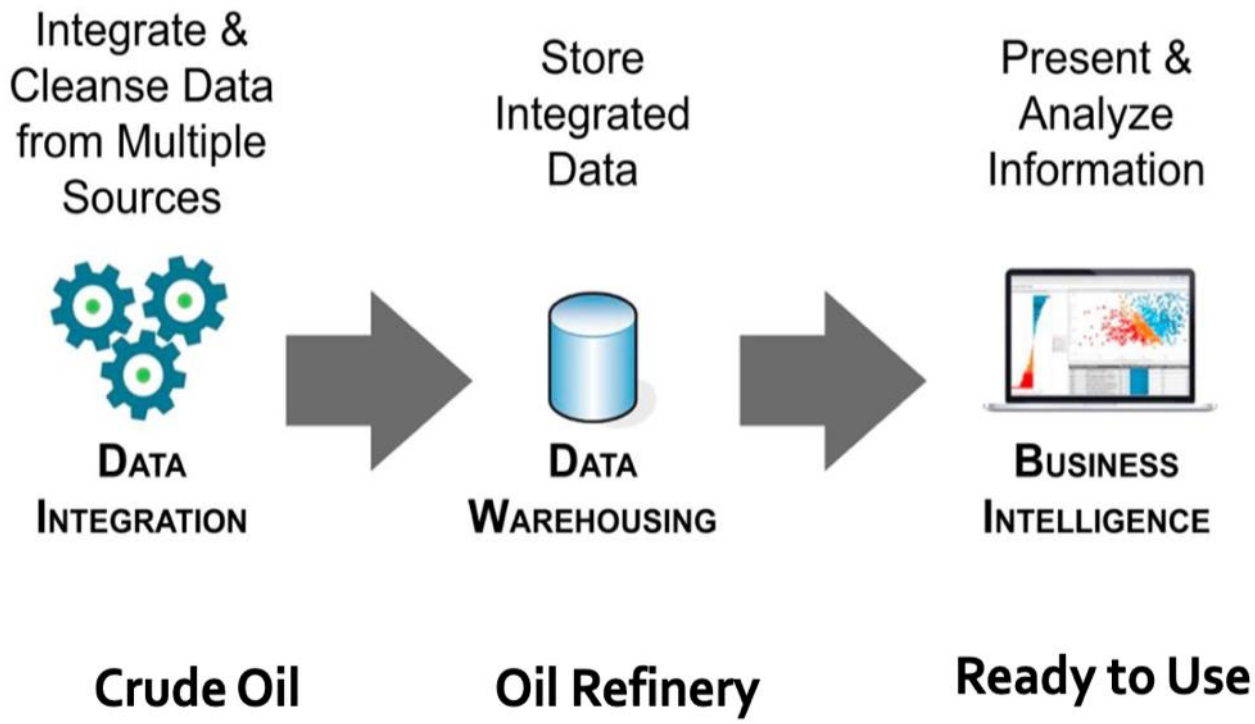
Operational / transactional systems – places where work is done and raw, siloed data is generated

Data integration – the process of aggregating and preparing data from multiple sources to be stored in a warehouse and present a unified view used by BI tools and applications

Data warehousing – the storing and/or staging of integrated information, separate from the operational systems, in optimized form for analysis

Business intelligence – tools and applications which present and analyze information

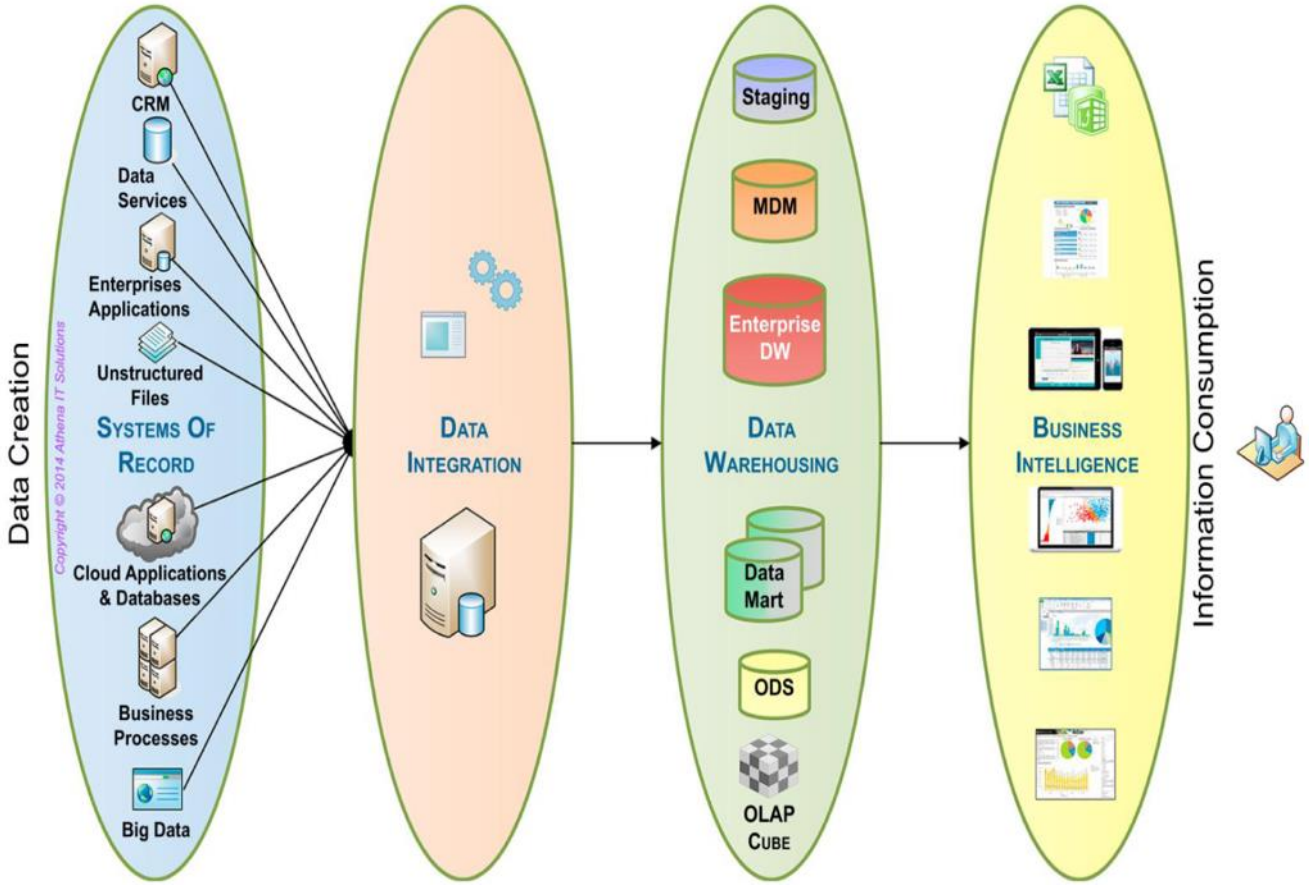
Big data - could mean really large, could mean unstructured, could mean overly complex, or could mean all of the above



* The 3 V's (volume, velocity and variety) are three defining properties or dimensions of big data. Volume refers to the amount of data, velocity refers to the speed of data processing, and variety refers to the number of types of data.

How Business Intelligence works

- Businesses and organizations have questions and goals. To answer these questions and track performance against these goals, they gather the necessary data, analyze it, and determine which actions to take to reach their goals.
- On the technical side, raw data is collected from business systems. Data is processed and then stored in data warehouses, the cloud, applications, and files. Once it's stored, users can access the data, starting the analysis process to answer business questions.
- BI platforms also offer data visualization tools, which convert data into charts or graphs, as well as presenting to any key stakeholders or decision-makers.

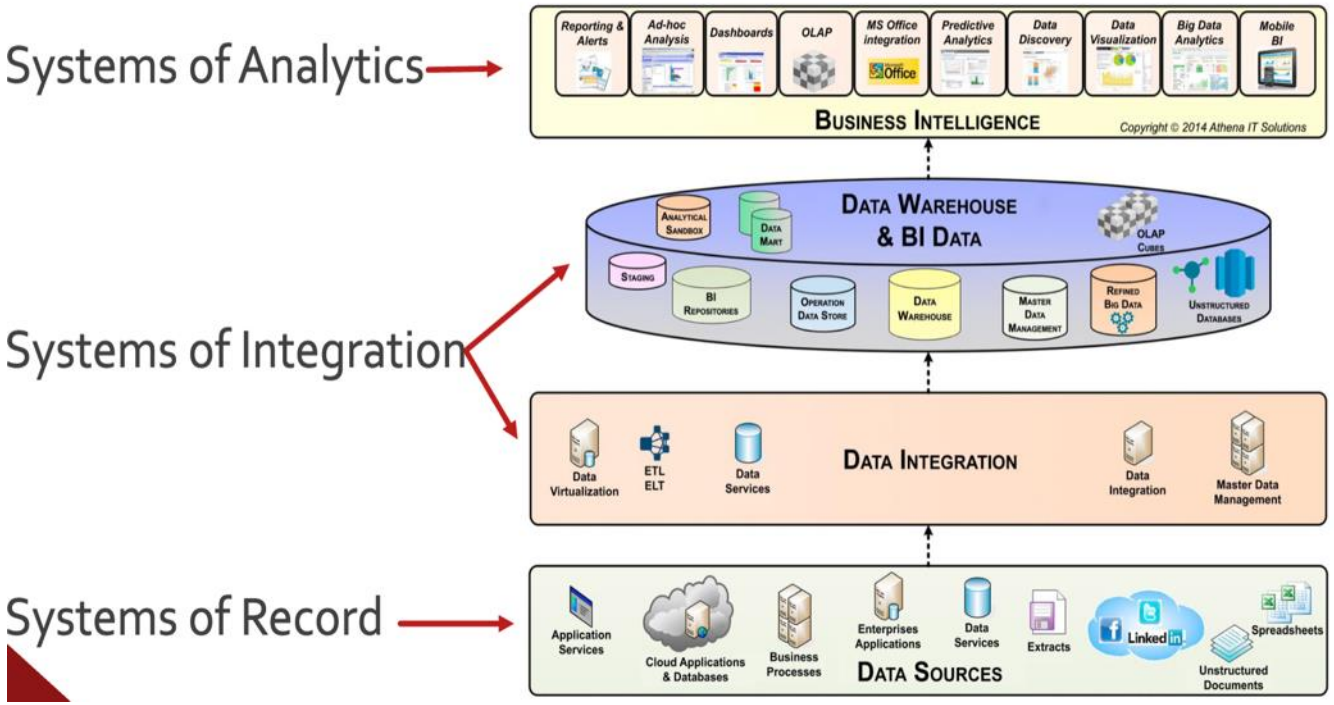


Different types of BI methods

- Business intelligence has evolved to include more processes and activities to help improve performance. These processes include:
  - Data mining: Using databases, statistics, and machine learning (ML) to uncover trends in large datasets
  - Reporting: Sharing data analysis to stakeholders so they can draw conclusions and make decisions
  - Performance metrics and benchmarking: Comparing current performance data to historical data to track performance against goals, typically using customized dashboards
  - Descriptive analytics: Using preliminary data analysis to find out what happened
  - Querying: Asking the data-specific questions, BI pulling the answers from the data sets
  - Statistical analysis: Taking the results from descriptive analytics and further exploring the data using statistics such as how this trend happened and why
  - Data visualization: Turning data analysis into visual representations such as charts, graphs, and histograms to more easily consume data
  - Visual analysis: Exploring data through visual storytelling to communicate insights on the fly and stay in the flow of analysis
  - Data preparation: Compiling multiple data sources, identifying the dimensions and measurements, and preparing it for data analysis

How BI, data analytics, and business analytics work together

- Business intelligence includes data analytics & business analytics and uses them as parts of the whole process.
- BI helps users draw conclusions from data analysis. Data scientists dig into the specifics of data, using advanced statistics and predictive analytics to discover patterns and forecast future patterns.
- Business analytics includes data mining, predictive analytics, applied analytics, and statistics." In short, organizations conduct business analytics as part of their larger business intelligence strategy



Traditional BI vs modern BI

Modern BI platforms provide continuous pulse monitoring of the organization through rapid analytics and empower businesses to accomplish mission objectives through predictive analytics.

Modern BI prioritizes self-service analytics and speed to insight. Traditional BI used a top-down approach where BI was managed by IT.

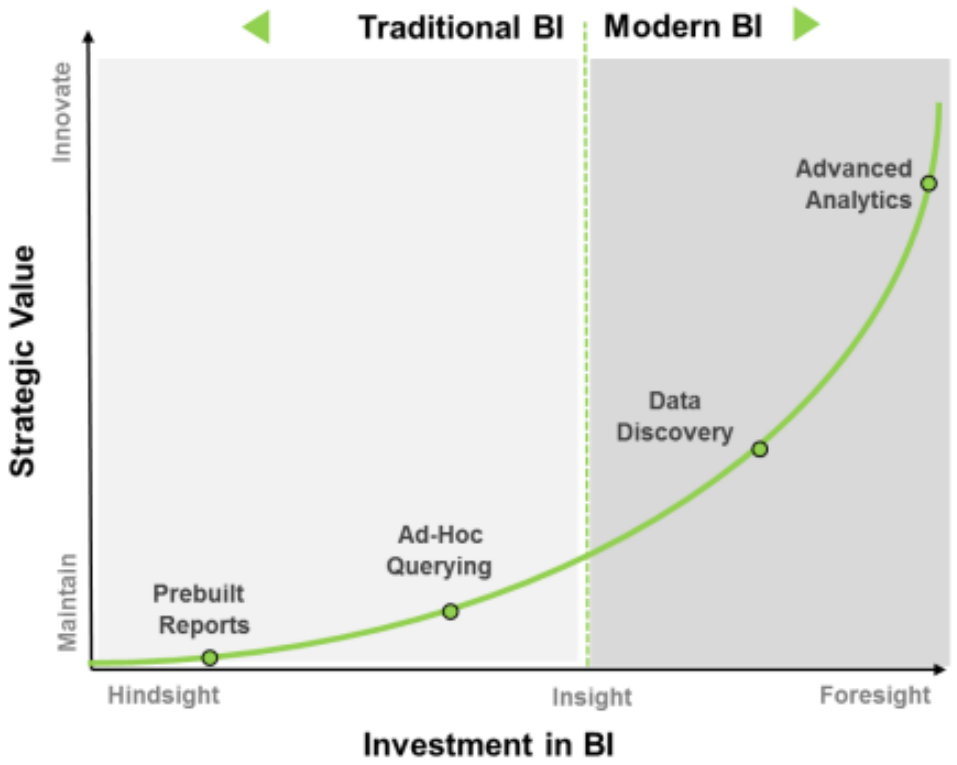


Figure 1: As organization investment in data modernization increases, value grows exponentially and changes from hindsight to insights to foresight.

Benefits of BI

- Great BI helps businesses and organizations ask and answer questions of their data.
- BI is more than just software—it's a way to keep a holistic and real-time view of all your relevant business data. Implementing BI offers a myriad of benefits, from better analysis to an increase in competitive advantage. ***Some of the top business intelligence benefits include:***
  - Data clarity
  - Increased efficiency
  - Better customer experience
  - Improved employee satisfaction
  - Reduced complexity
  - Trend, insights, patterns & relationships
  - Improved data access & lower costs
  - Right time, right data

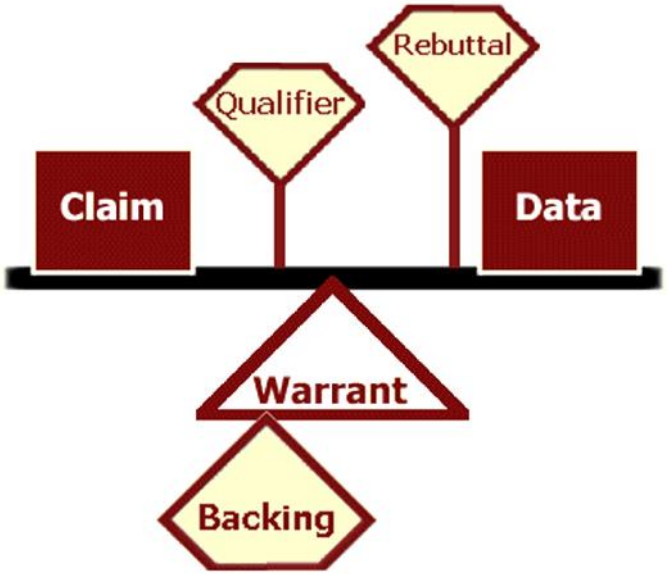
BI Dashboard

- Business intelligence dashboards are information management & data visualization solutions to analyze data.
- Content creators can use interactive elements like filters and actions to combine charts, graphs and reports in a single screen for snapshot overviews.
- Dashboards are one of the most popular capabilities of BI platforms because they present easily understandable data analysis, allow you to customize which information you want to view, and provide a way to share the results of your analysis with others.
- Dashboards and reports are critical for business intelligence, and dashboards can help users understand complex reports. Dashboards are ideal for stakeholders who need at-a-glance overviews of performance. Reports are for stakeholders who need more details and who want to slice and dice data to uncover insights.

https://www.datacamp.com/blog/9-power-bi-dashboard-examples

Claim-Data-Warrants- A Model for Analyzing Arguments

- A claim is something the author wants the audience to believe. Data is the evidence or appeal they use to convince the reader to believe the claim. A warrant is the (often implicit) assumption that links the data with the claim.
- Stephen Toulmin identifies three essential parts of any argument as the claim; the data (also called grounds or evidence), which support the claim; and the warrant. The warrant is the assumption on which the claim and the evidence depend. Warrant explains why the data support the claim



Claim: You should buy our tooth-whitening product.

Data or Grounds: Studies show that teeth are 50% whiter after using the product for a specified time.

Warrant: People want whiter teeth.

Data Visualization Redesign – LinkedIn top skills dataset

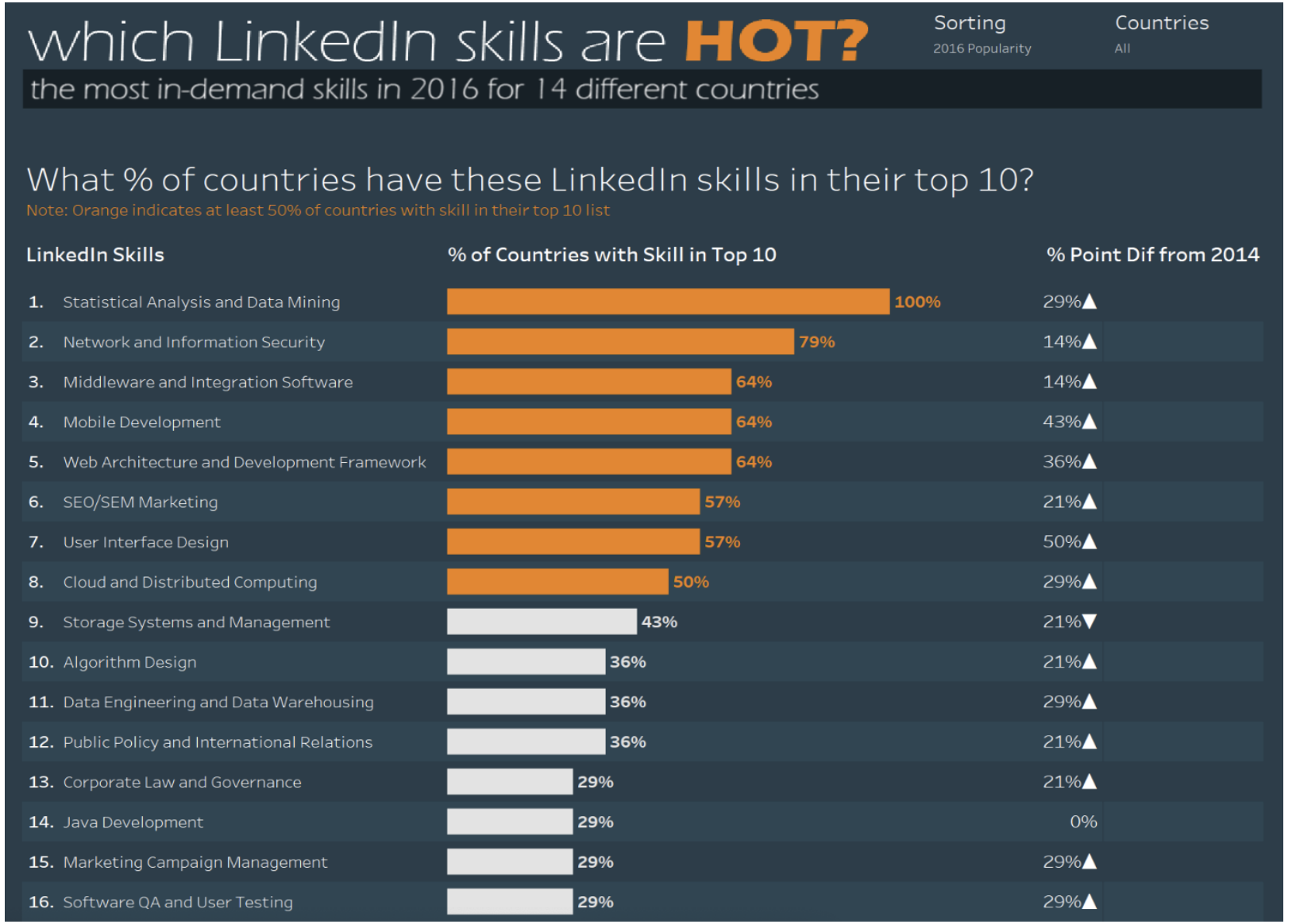
- The objective of this project is to redesign a data visualization (e.g., an infographic, a data visualization in a newspaper, a project report, or a sales report). This will allow you to learn from others, sharpen your critical perspective on data visualizations, reason about design decisions, and attempt to improve data visualizations.

- **Dataset description:**
  - LinkedIn top skills dataset identifies skills which are in demand based on their respective assigned ranks. The skills are ranked differently in every country and some countries prioritize a certain skill more or less as compared to other nations. The dataset also categorizes the skills on yearly basis and the period from 2014 to 2017 inclusively is used as a framework for analyzing the top skills.
  - The dataset has four columns Skill, Country, Year and Rank and has over 800 rows of records

- The most important objective to prove is the claim that statistical analysis, data analytics and cloud computing are becoming top ranked LinkedIn skills and therefore their demand is increasing. The ranking range used in the dataset is from 1 to 25 where 1 indicates the highest ranked skill.

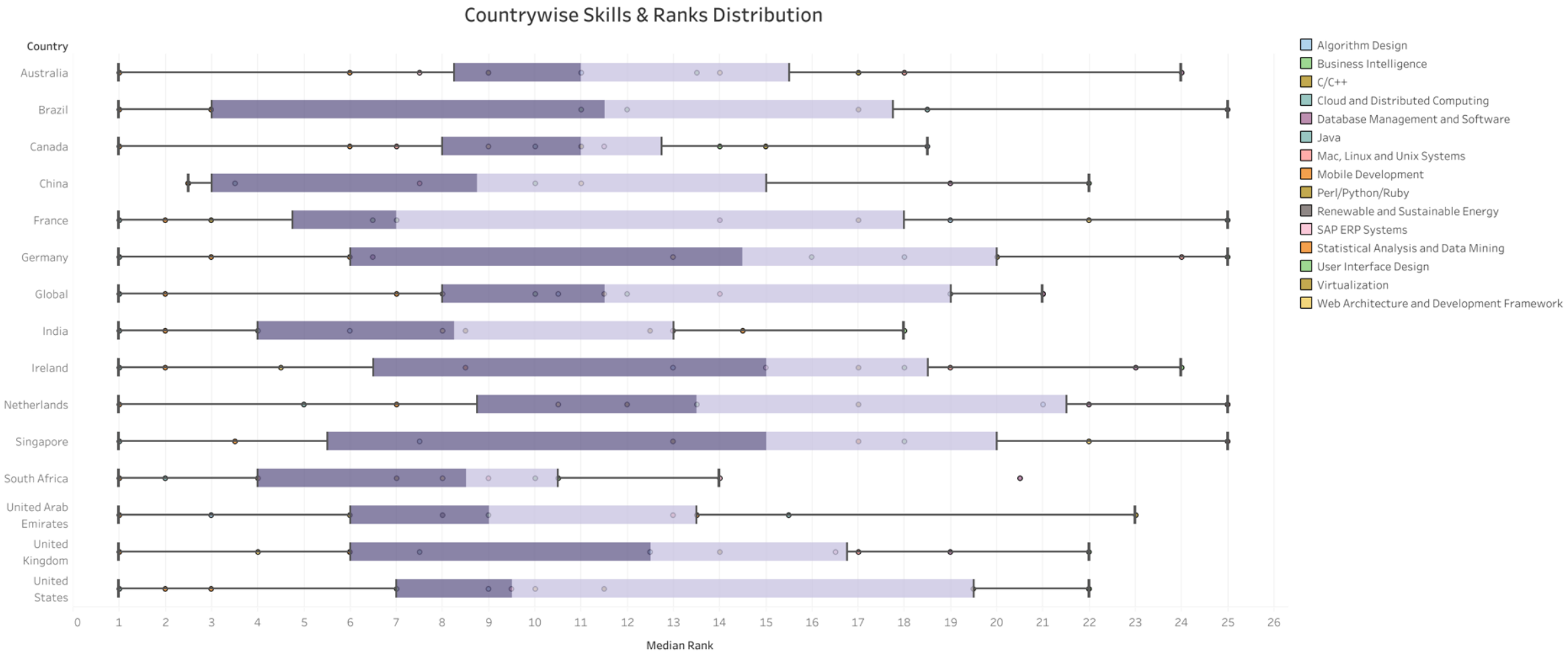
- https://github.com/sultanadeel/Redesign_Project

LinkedIn top skills



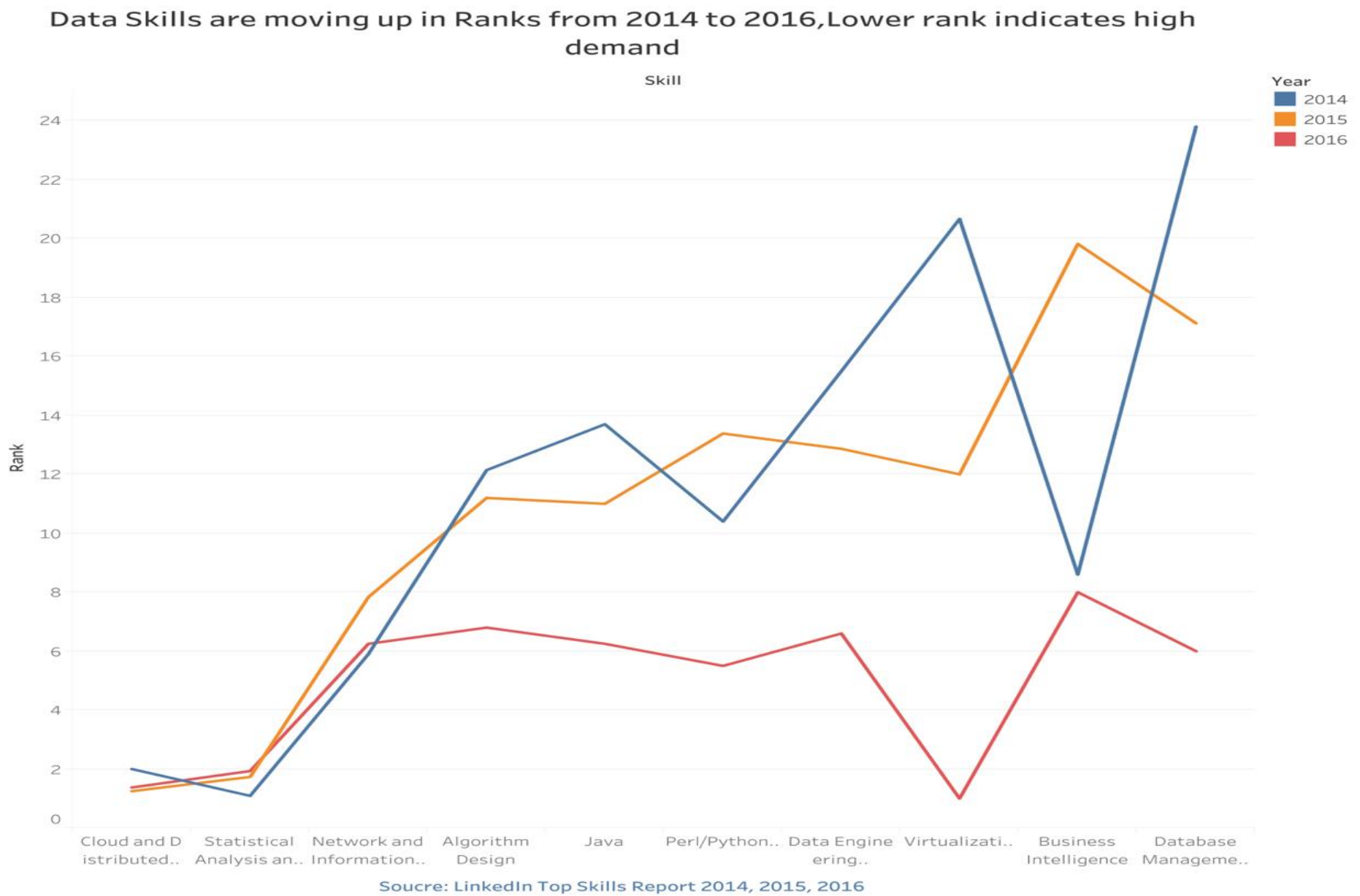
data source: LinkedIn's annual Top Skills Reports 2014, 2015, and 2016

Claim: Top Ranked skills are present in all the major economies as well as the emerging ones

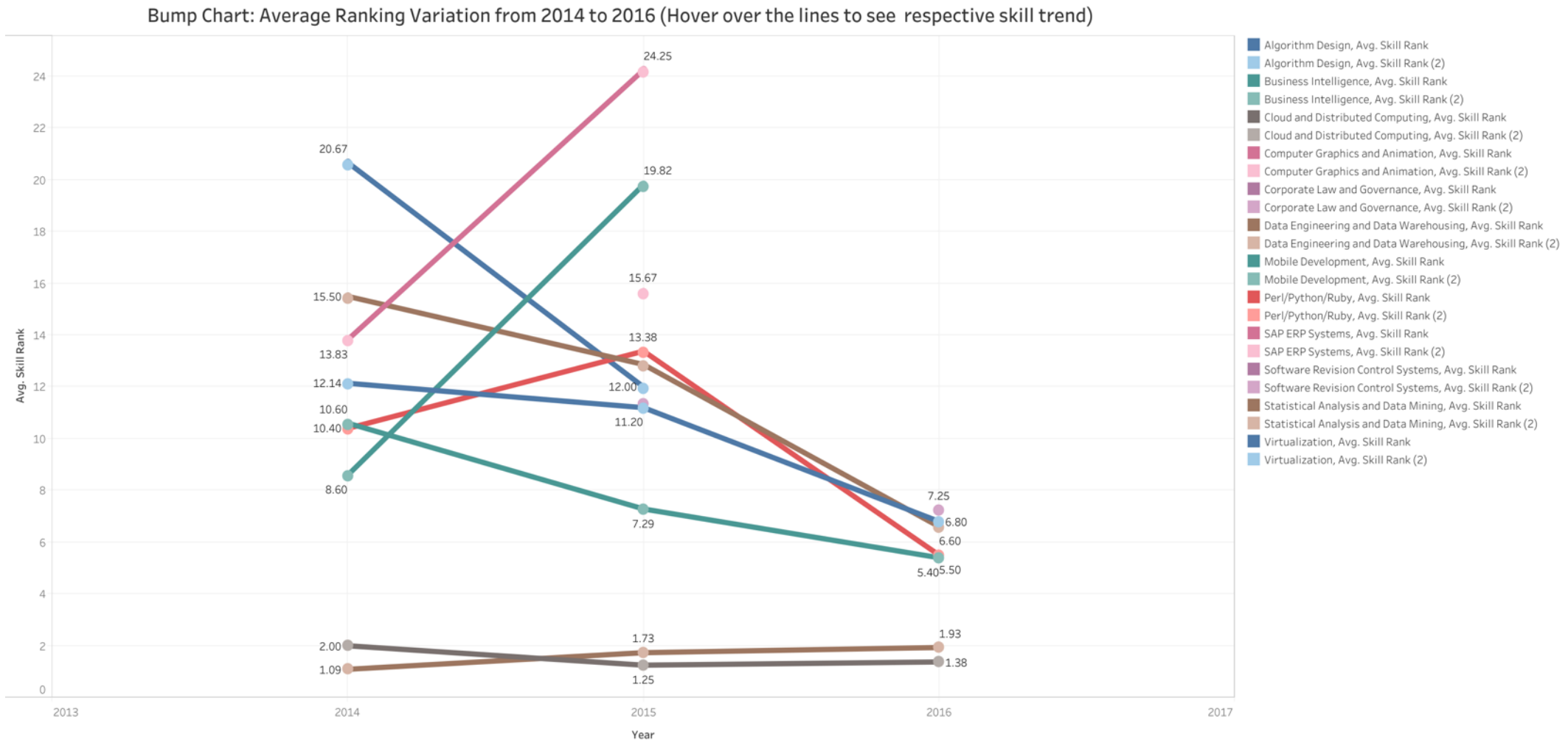


Soucre: LinkedIn Top Skills Report 2014, 2015, 2016

Claim: Some technologies become popular simultaneously in all the countries, while other skills emerge from major economies and then spread across the globe

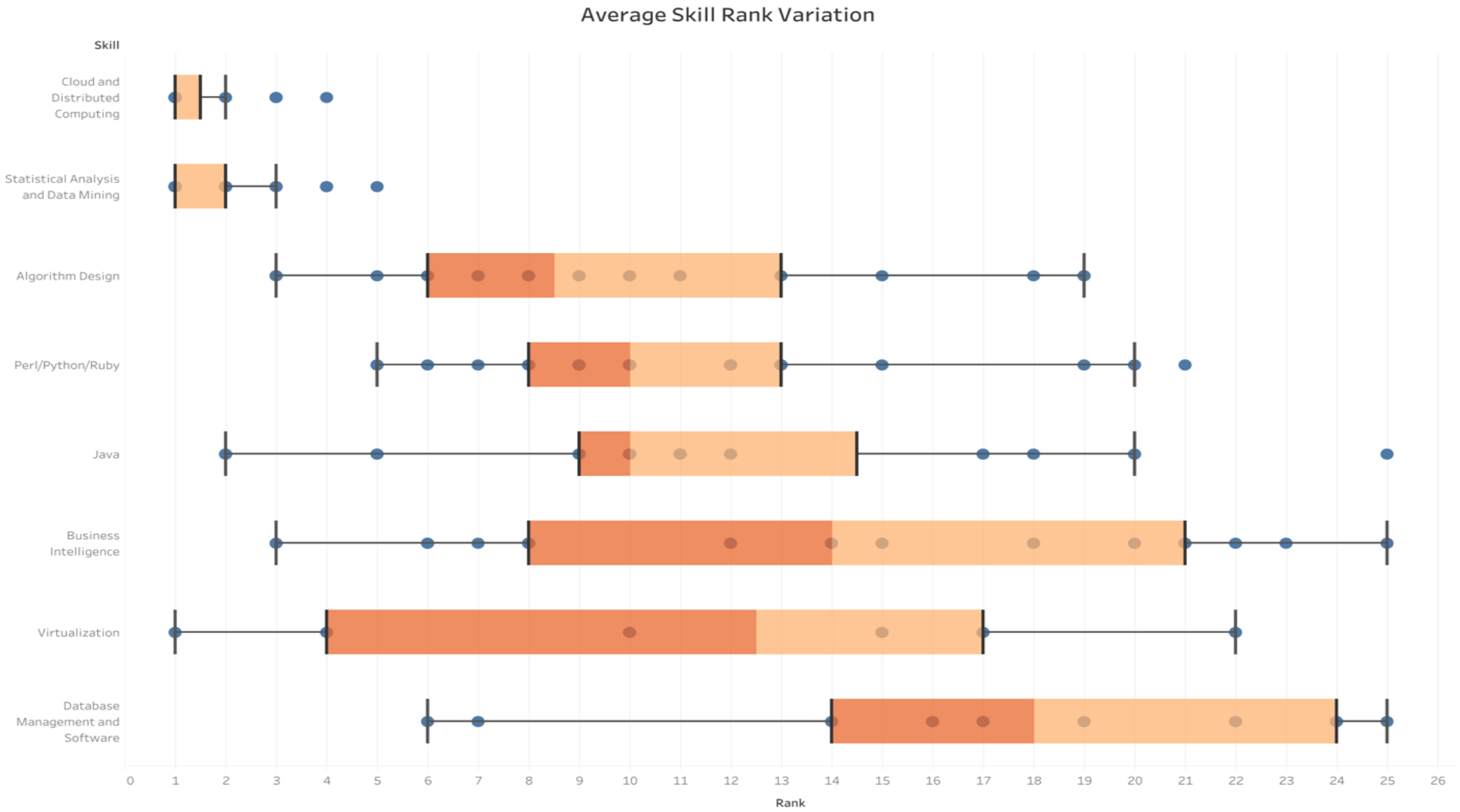


Claim: Increasing popularity in top skills is encouraging the previously less popular skills to become more demanding and rewarding



Soucre: LinkedIn Top Skills Report 2014, 2015, 2016

Claim: Cloud & Statistical Data Mining skills are consistently ranked higher among all the countries as shown by the positive skewness of their respective boxplots

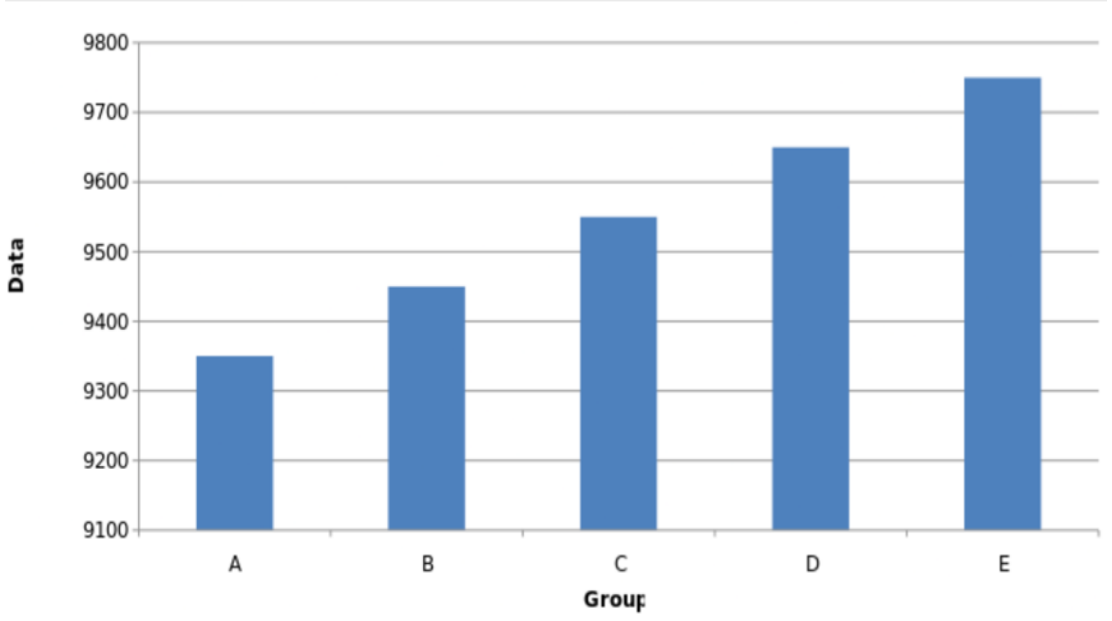


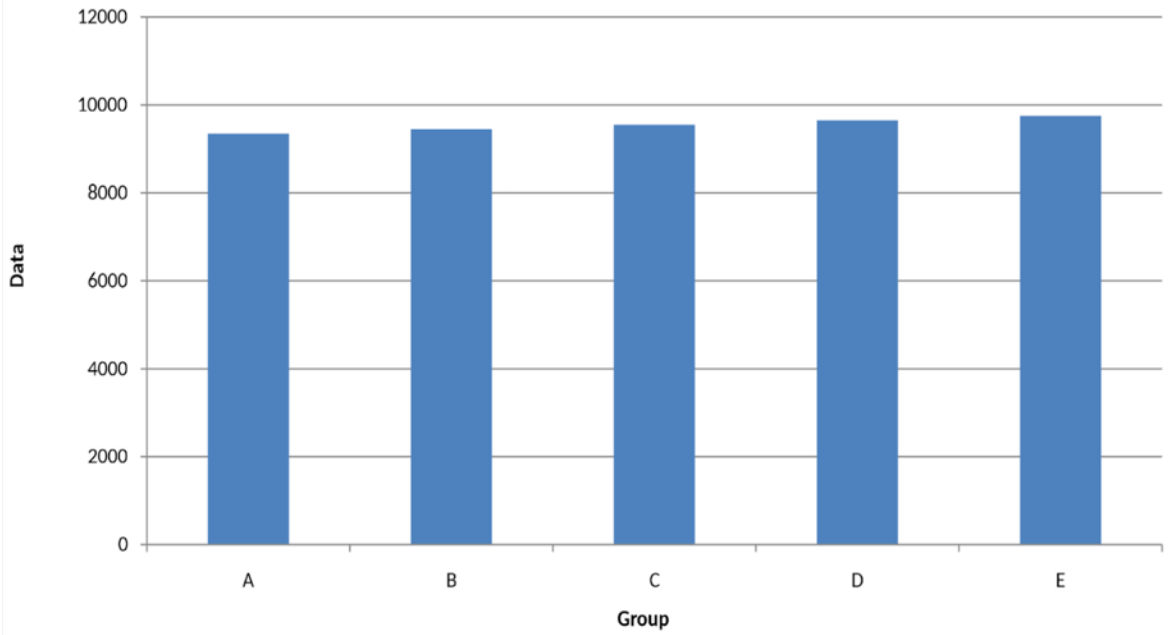
Soucre: LinkedIn Top Skills Report 2014, 2015, 2016

Deceptive Visualization

- The objective of this project is to create a data visualization that purposefully distorts the data or deceives the reader. You start with an existing data visualization and develop a deceptive version of it. You may not alter the data. The purpose of this project is for you to experience and realize the ethical implications of design decisions during the development of data visualizations.

- **<u>Truncating y-axis in graphs:</u>** A very common misleading data visualization example is changing the value of the y-axis's starting point from zero to any other number. This blows up the differences when comparing data:

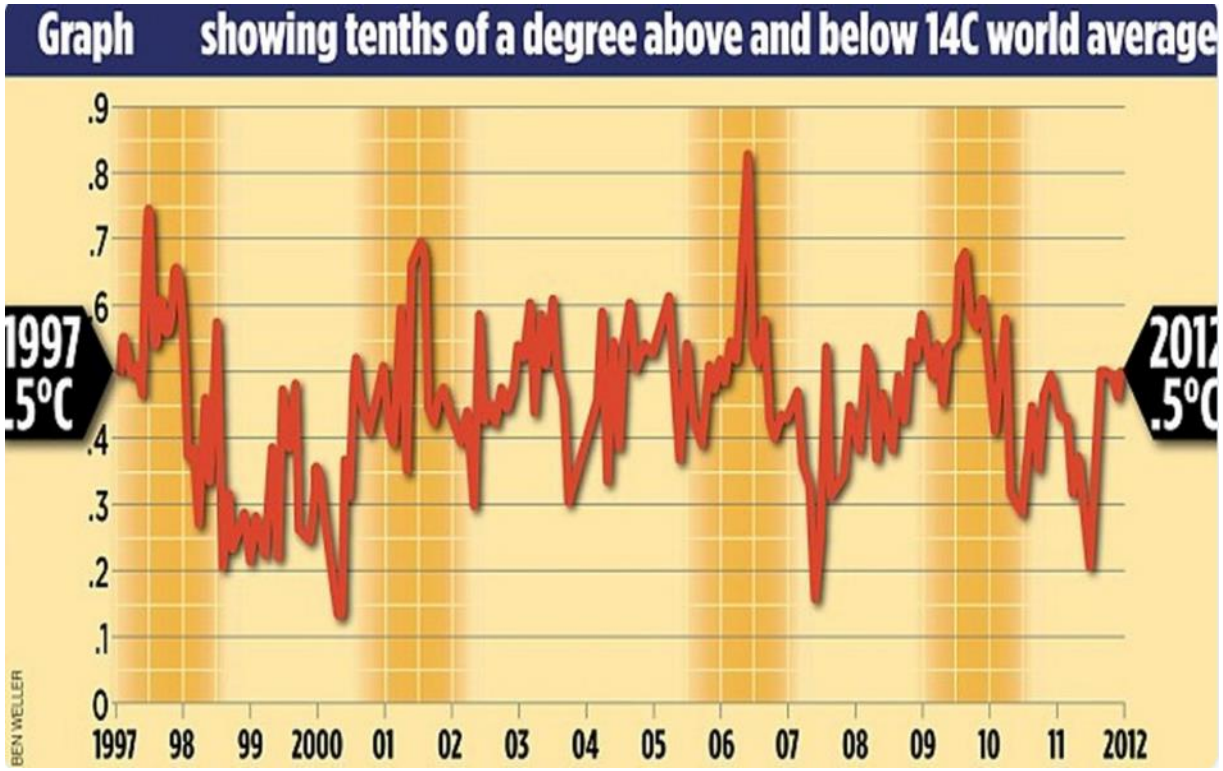


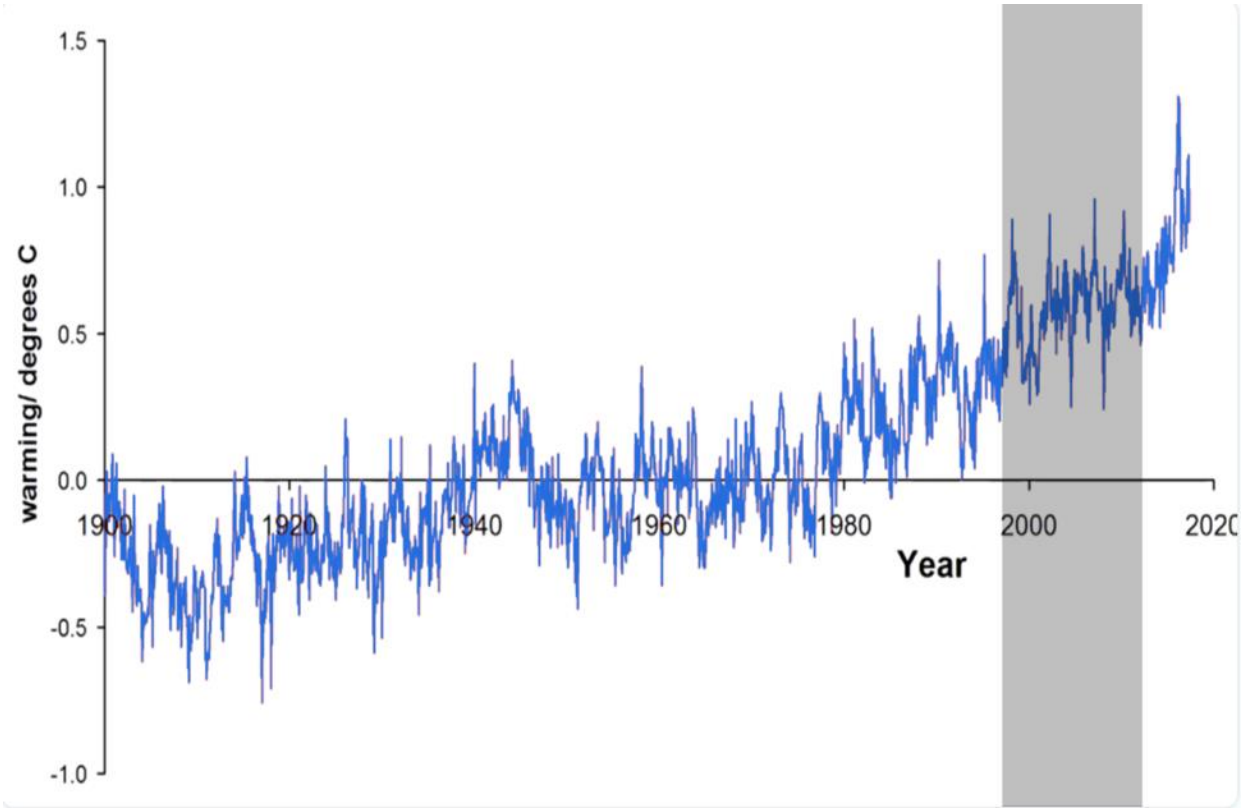


* From the graph left, all data is in the 9300-9800 range on the y-axis and this tends to show an exaggerated growth. This data can be much more accurately presented in the right graph. Now we can see that the differences between the data of the different groups are small. This is a more accurate picture of the differences and a fairway to present the information.

Deceptive Visualization

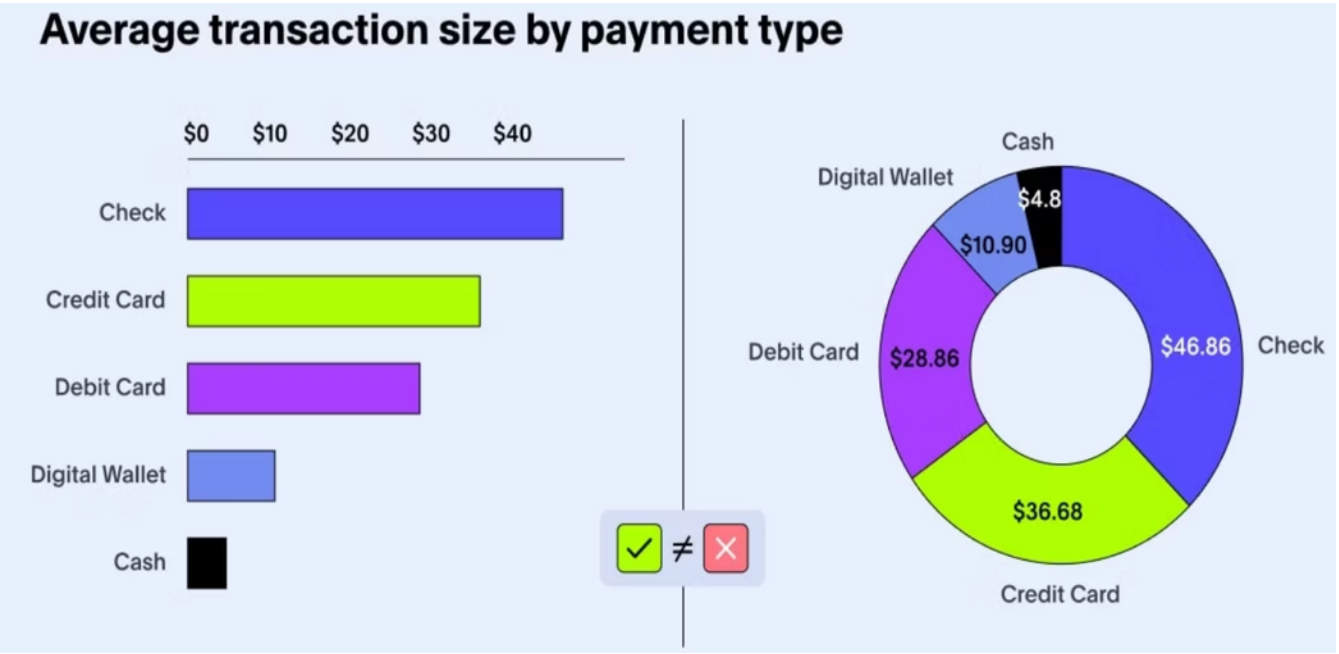
Improper extraction – Cherry Picking: The graph below (left) is a good misleading data visualization example. This graph shows the average global temperature from 1997 to 2012. The goal of this graph is to prove that global warming is not happening by not looking at the previous data. The chart on the right shows the average temperatures from 1900 to 2020, showing a clear increase in global temperatures.

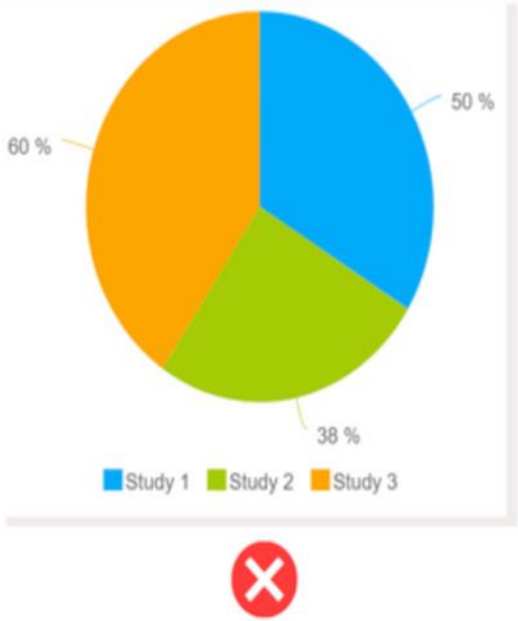


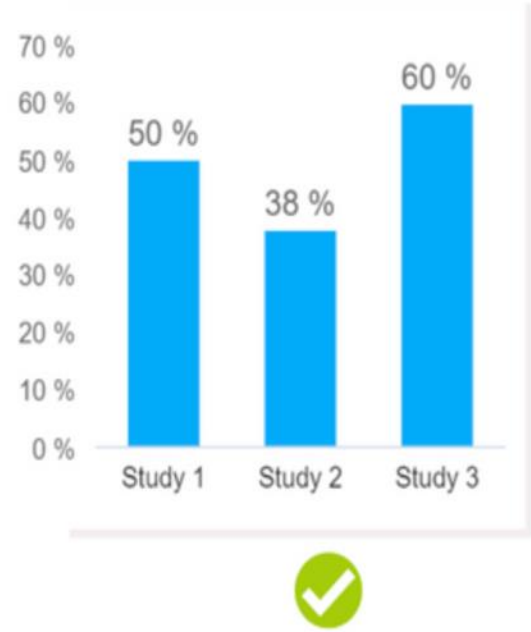


Deceptive Visualization

- **<u>Using wrong chart type:</u>**
  - ➢ Take into account your stakeholder preferences, expertise levels, and communication styles.
  - ➢ Study the form and function of each chart type.
  - ➢ Ask yourself whether the chart or graph serves a distinct purpose.
  - ➢ Think about what message you want to convey.







* the pie chart is a poor choice as the visuals don't represent the numbers in their best or most accurate format. Typically, a pie chart should amount to a whole (100 expressed as a percentage) – but that's not the case in this instance. As you can see, a neat, balanced bar chart is the best choice as it presents the information in its truest, most digestible form.

Australia Gender Pay Gap

- Our main claim is that “Australia’s top 50 jobs pay Men more than Women”. We analyzed that top paying jobs such as Neurosurgery, Ophthalmology and Internal Medicine pay men more than woman even though women are competitively qualified and capable to perform equally well.

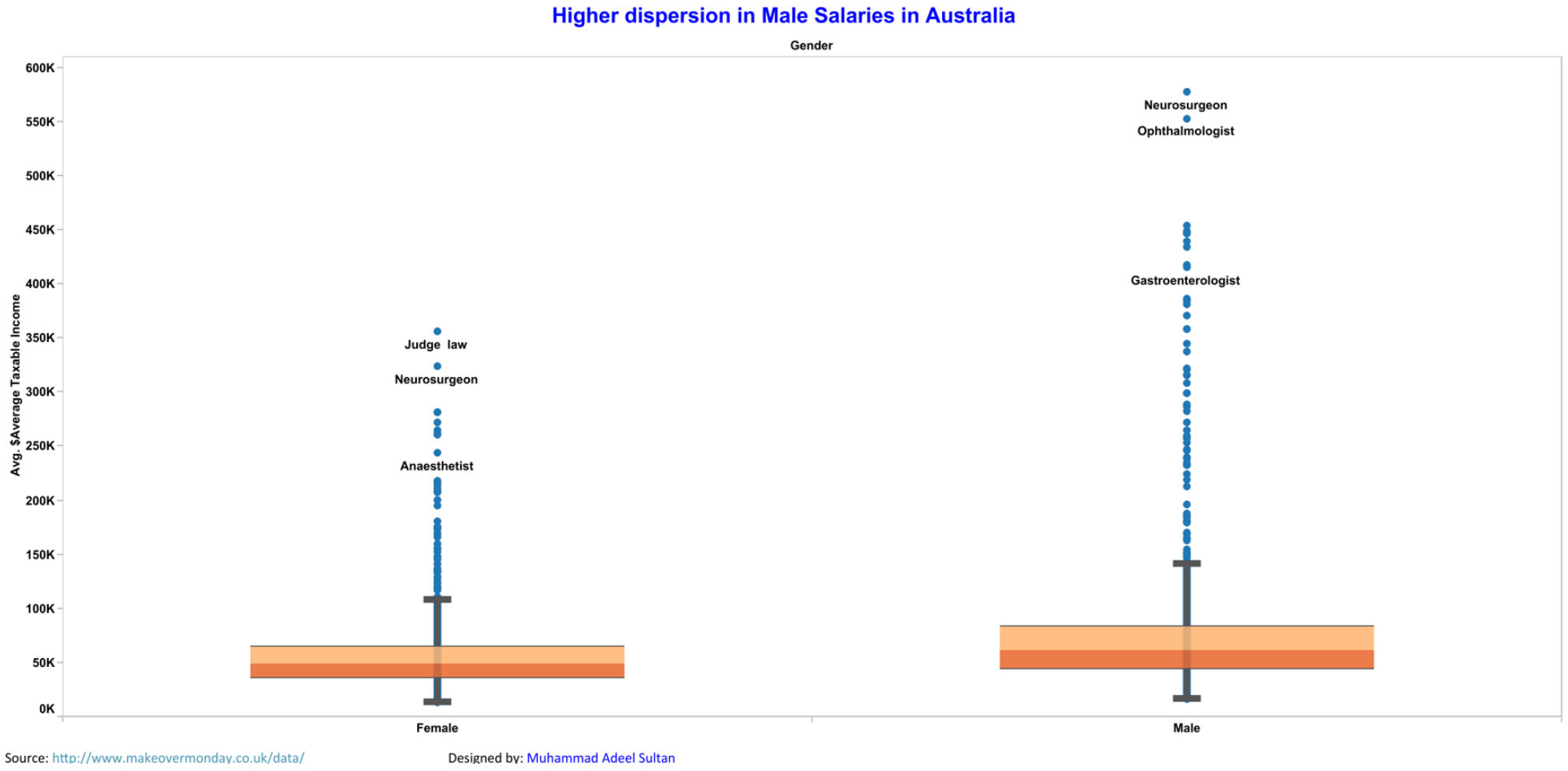
- Dataset description:

  - The dataset is provided by the Australian government for over 1,000 related occupations and the respective salaries for Men and Women under each occupation. The occupations are ranked from 1, Neurosurgery being the highest paid profession and 1104 being the lowest paid. The dataset has 5 columns namely, Gender Rank; which ranks every profession differently for Men and Women, Occupation name, Gender, Number of Individuals in each profession, and average taxable Income with the occupation and gender.

- This project is aimed at creating two contradictory Tableau Visualizations using a selected dataset on a contentious and debatable topic. Initially, we will develop a visualization that conforms to our main claim and then by incorporating external data in the existing dataset using Python, we will create a rebuttal Tableau visualization that would refute our original claim by providing a legitimate and valid arguments, claims and warrants

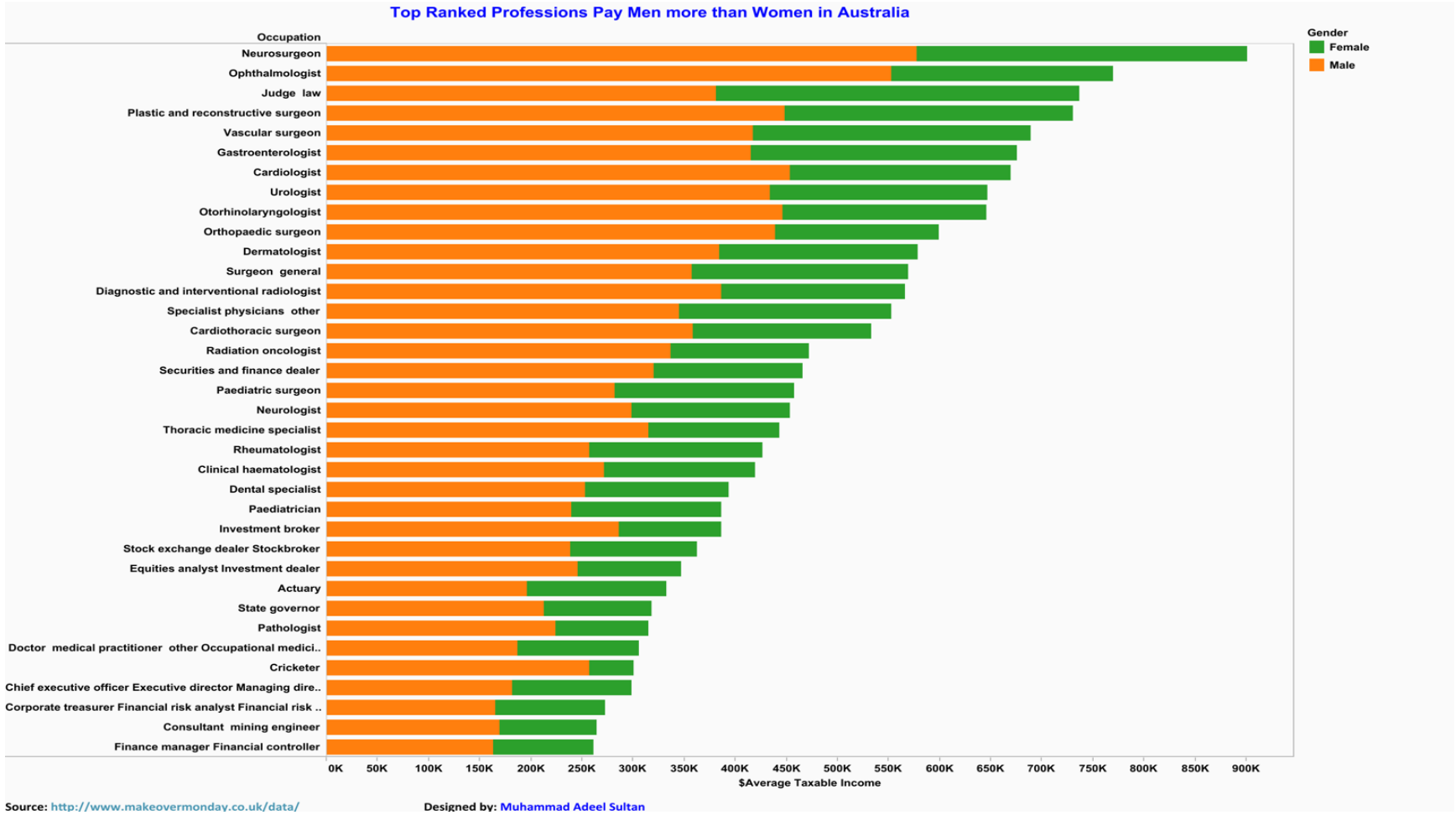
- https://github.com/sultanadeel/Australia-s_Gender_PayGap

Claim: Male salaries in top professions in Australia are highly dispersed



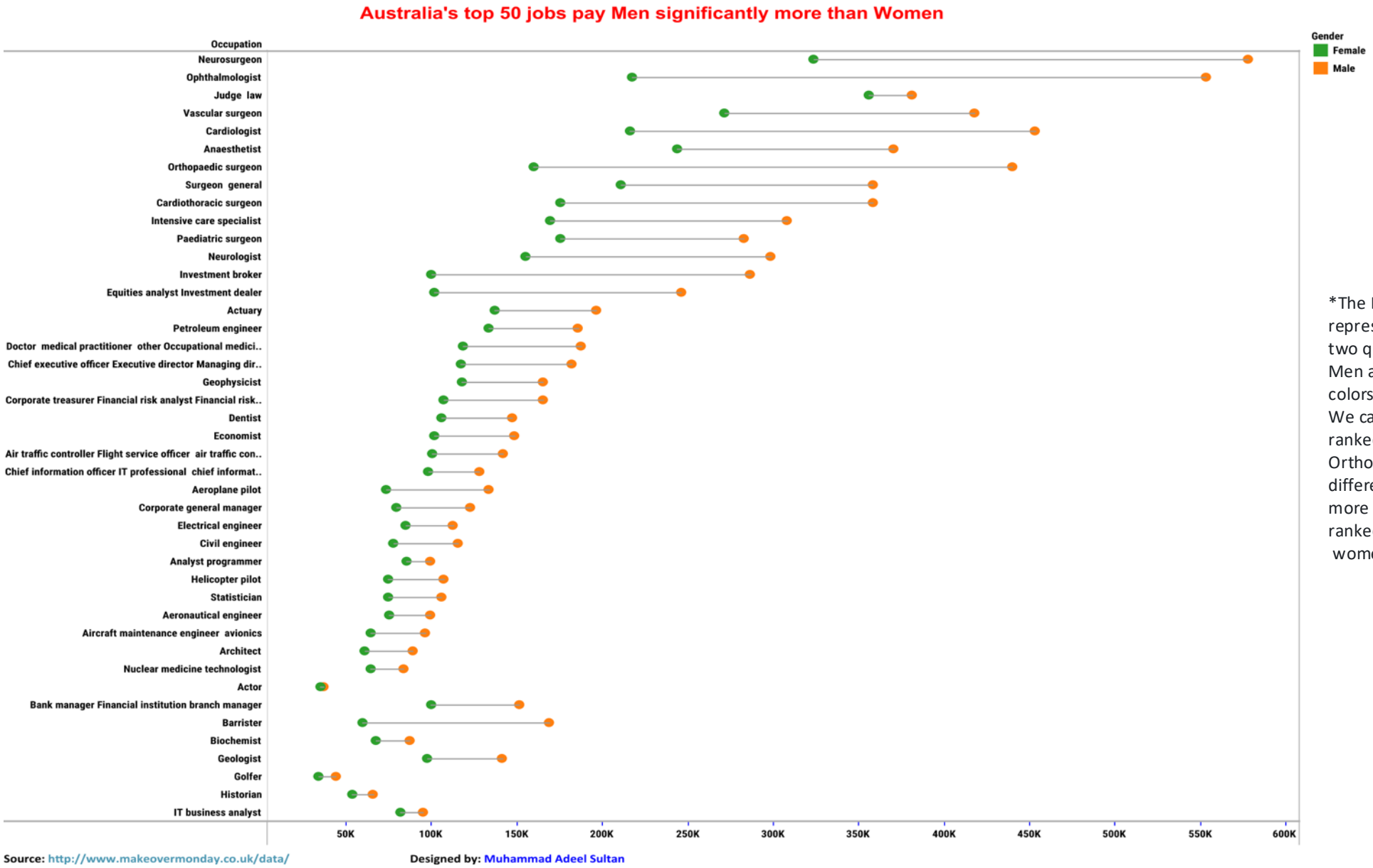
*the median of the boxplot for Men is relatively higher than Women and the Men boxplot is skewed to the right and has positive skewness. The outliers for Men boxplot are comparatively greater than for Women boxplot, indicating the fact most of the top occupations are paying Men more than Women in Australia for specific reasons that we would analyze in the rebuttal section of this analysis. We can also see in the boxplot visual that for some high ranked professions like Neurosurgery, there is a wide gap between the Men and Women Average Taxable Income. Men are commanding around $550,000 vs Women only $320,000 for a Neurosurgeon occupation.

Claim: Top Ranked professions pay men more than women in Australia



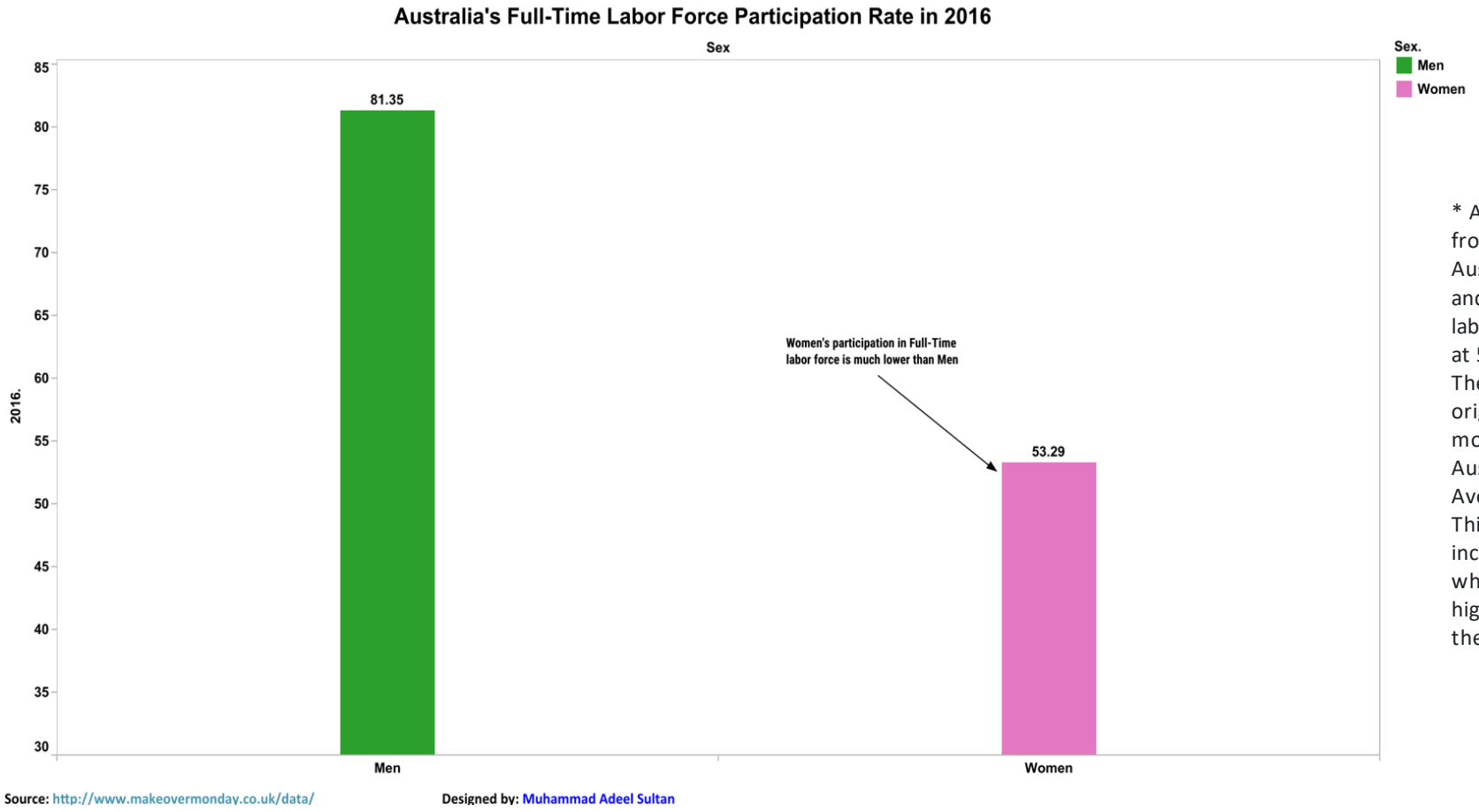
*The visual shows that as the rank of occupation gets lower, the Average Taxable Income variation between Men and Women gets narrower. Therefore, the top ranked professions such as Neurosurgery, Ophthalmology, Judge Law, Plastic Surgery and Cardiology are some of the professions which pay Men significantly more than Women.

Claim: Australia's top 50 jobs pay Men significantly more than Women



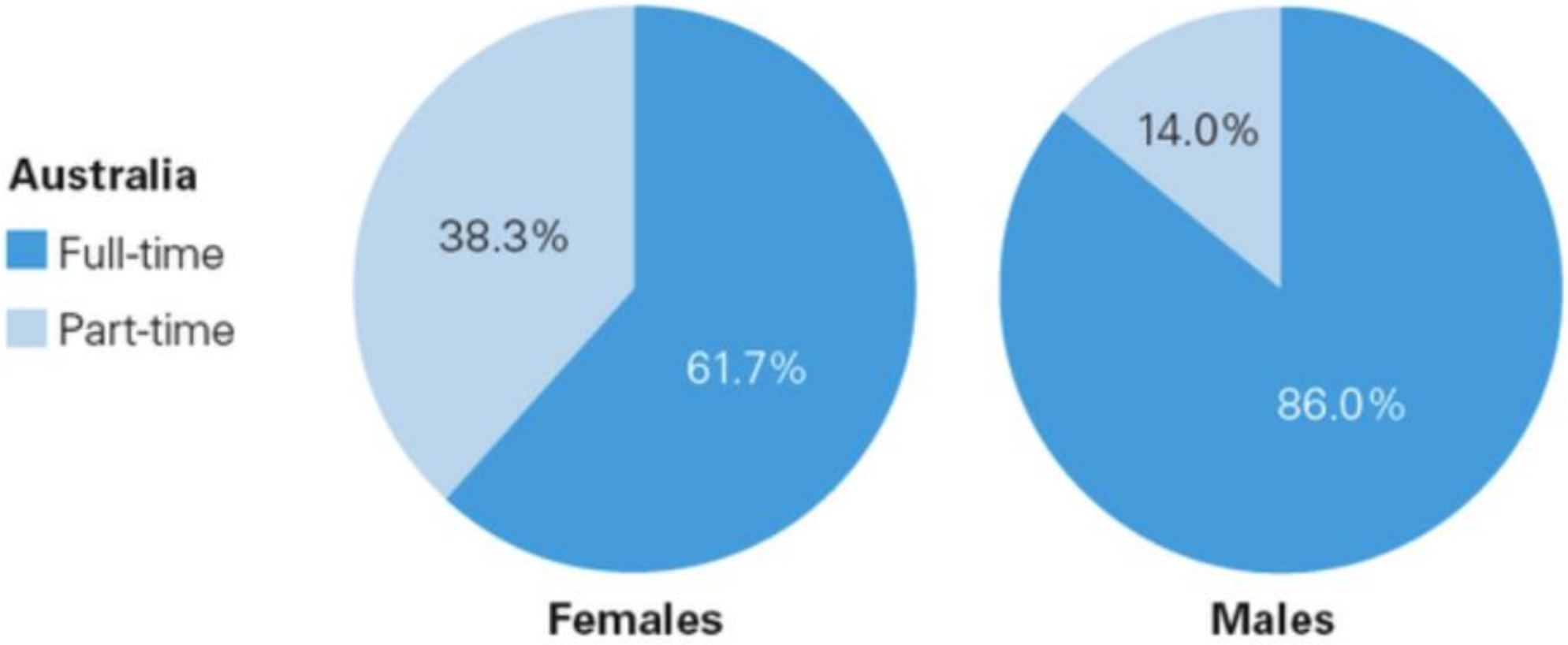
*The DNA chart is an accurate reflection of this claim, it represents the actual difference or spread between the two quantities which are Average Taxable Income of Men and Women represented by blue and yellow colors respectively.
We can see that the biggest difference is in the top ranked profession of Neurosurgery, Ophthalmology, Orthopedic Surgery and Cardiology. The bigger the difference in the two quantities in the DNA chart, the more persuasive our claim would be that the top ranked jobs are paying men more than women.

Rebuttal to the original claim: Full Time Labor force participation of Women is less than Men in Australia



* After incorporating external data from http://www.oecd.org/gender/data/employment/ on Australia's total and full time labor force participation by men and women, we analyzed that women's participation in full time labor force in Australia is significantly less than Men and stands at 53.3% against 81.4% for Men in the year 2016
Therefore, our Rebuttal claim which contradicts and refutes our original claim that "Australia's Top 50 jobs pay Men significantly more than Women" is that "Since Women's participation in Australia's Full time labor force is significantly less than Men, the Average Taxable Income of Women is less than Men".
This analogy is focused more on the top 50 occupations including Neurosurgery and Medicine related occupations where full time availability is of utmost importance and of higher demand by the employer considering the risk involved in the occupation

Rebuttal to the original claim: Full Time Labor force participation of Women is less than Men in Australia



https://www.ussc.edu.au/analysis/women-at-work-australia-and-the-united-states

Investing in cryptocurrencies

- Our targeted audience is a potential investor interested in diversifying portfolio of investments. In order to maximize expected return and minimize foreseeable risk , the ultimate objective is to invest in a industry different than the existing ones, and we are analyzing the positive and negative aspects of investing in cryptocurrencies, specifically Bitcoin

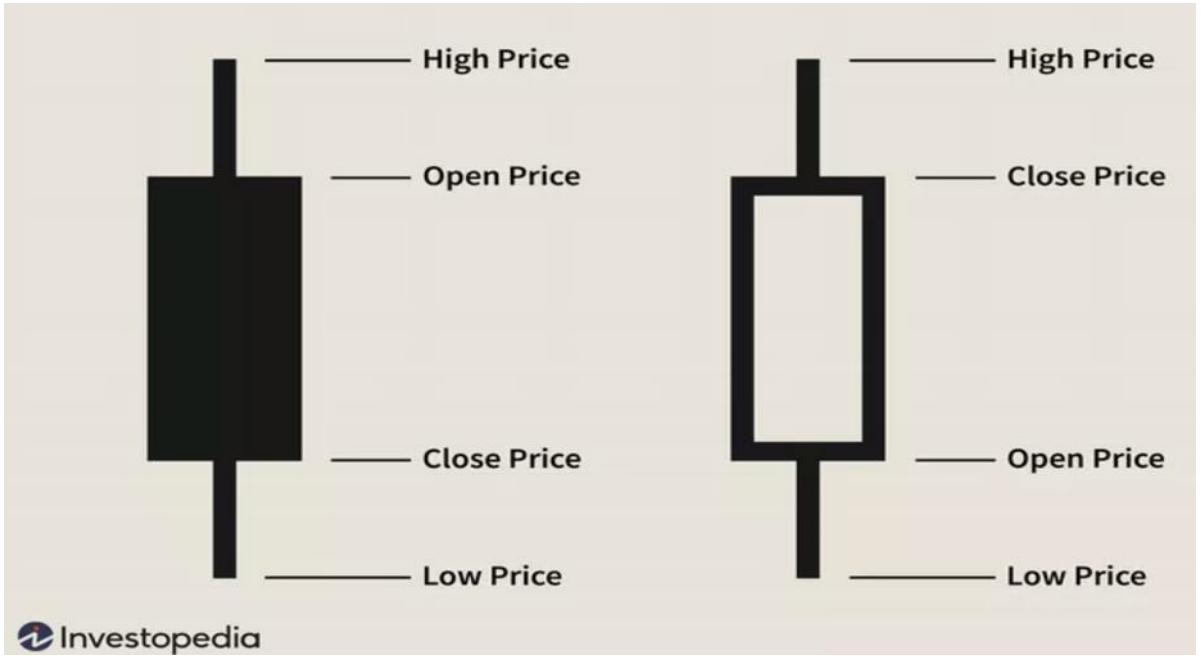
- **Audience Need:**
  - The Audience is interested in taking profits from the likely arbitrage opportunity embedded in cryptocurrencies. As we see in the patterns and trends of cryptocurrencies historical prices, there are short-term spikes and upticks in the prices which has profit making opportunity for prospective investors.

- **Audience Want:**
  - Our audience for this analysis of Bitcoin investment is a prospective investor intending to diversify investments.

- **Audience Fears:**
  - Cryptocurrencies are not backed by any government or a regulatory agency and the operations are decentralized as compared to the traditional centralized banking system and financial markets. Therefore, a general perception and an expected fear of our audience would be the safety, reliability and foreseeability of their investment in cryptocurrencies
  - https://github.com/sultanadeel/Team-project---The-Bitcoin

Candlestick Charts

- A **candlestick chart** is a style of financial chart used to describe price movements of a security, derivative, or currency.
- While similar in appearance to a bar chart, each candlestick represents four important pieces of information for that day: open and close in the thick body, and high and low in the "candle wick". Being densely packed with information, it tends to represent trading patterns over short periods of time, often a few days or a few trading sessions
- Candlestick charts also used for iot sensor data or timeseries data
- **<u>Candlestick Components</u>**
  - Just like a bar chart, a daily candlestick shows the market's open, high, low, and <u>close</u> prices for the day. The candlestick has a wide part called the "real body."
  - This real body represents the price range between the open and close of that day's trading. When the real body is filled in or black (also red), it means the close was lower than the open. If the real body is white (or green), it means the close was higher than the open.

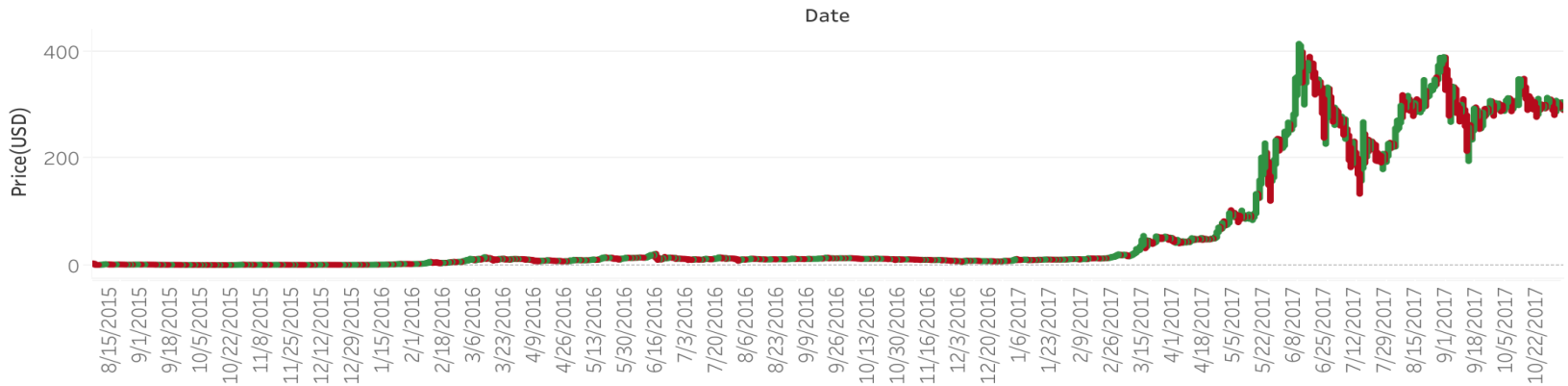


Bitcoin vs Ethereum

Bitcoin, less volatile than Ethereum!

Bitcoin's Candlestick!

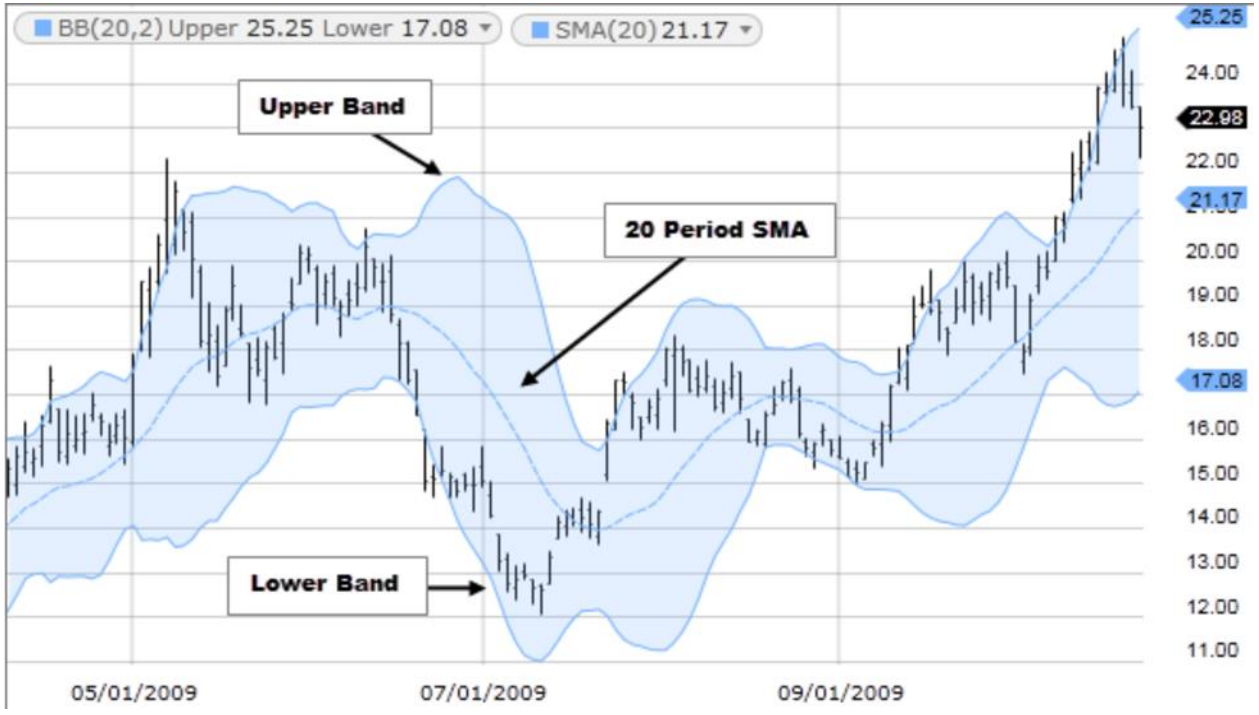
Ethereum Candlestick!



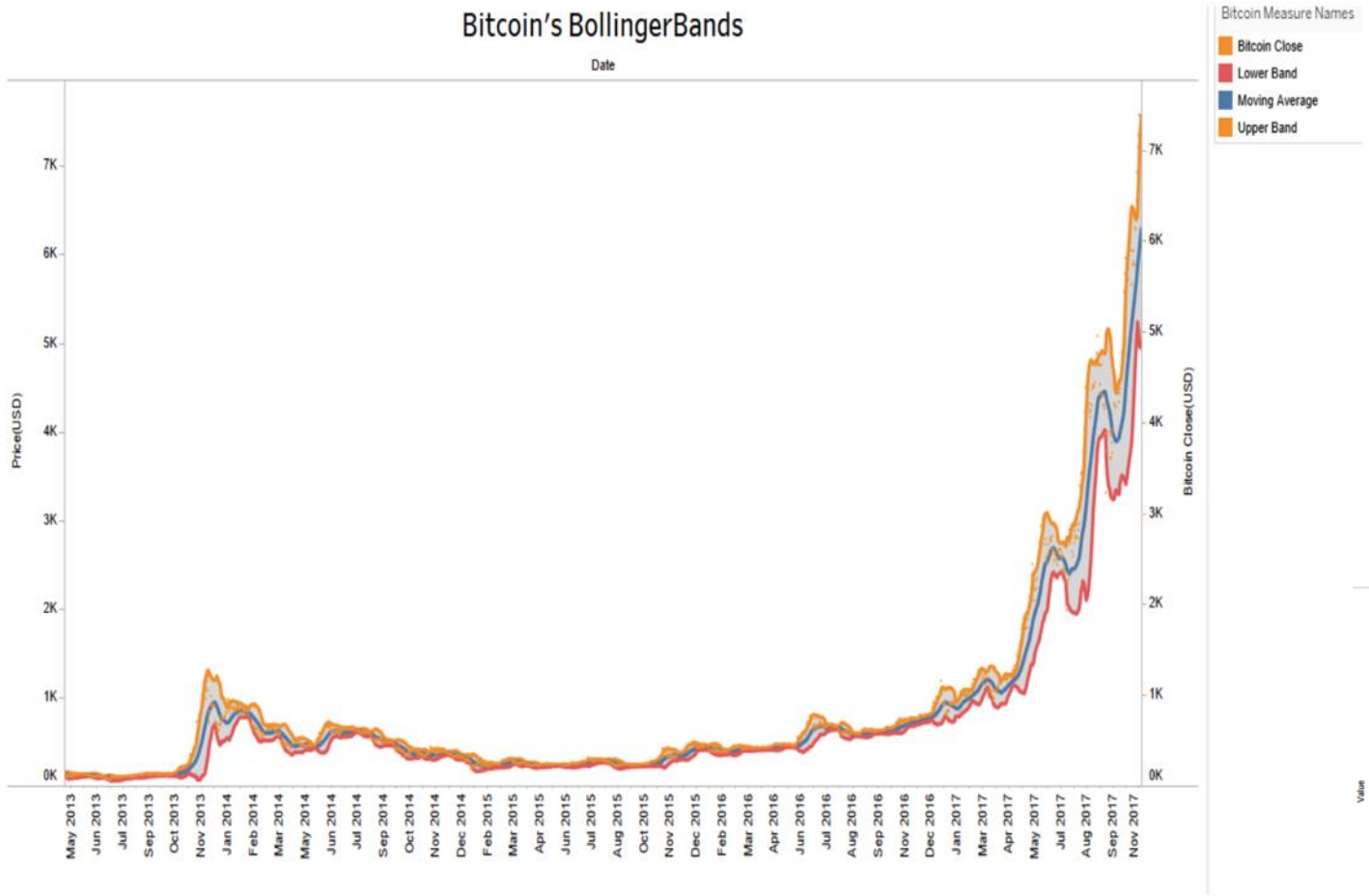
*Bitcoin has witnessed a tremendous increase in its market price since inception and has remained the least volatile cryptocurrency in terms of investment and the price appreciation over the recent years when compared to Ethereum, which had a wide volatility rate in 2017

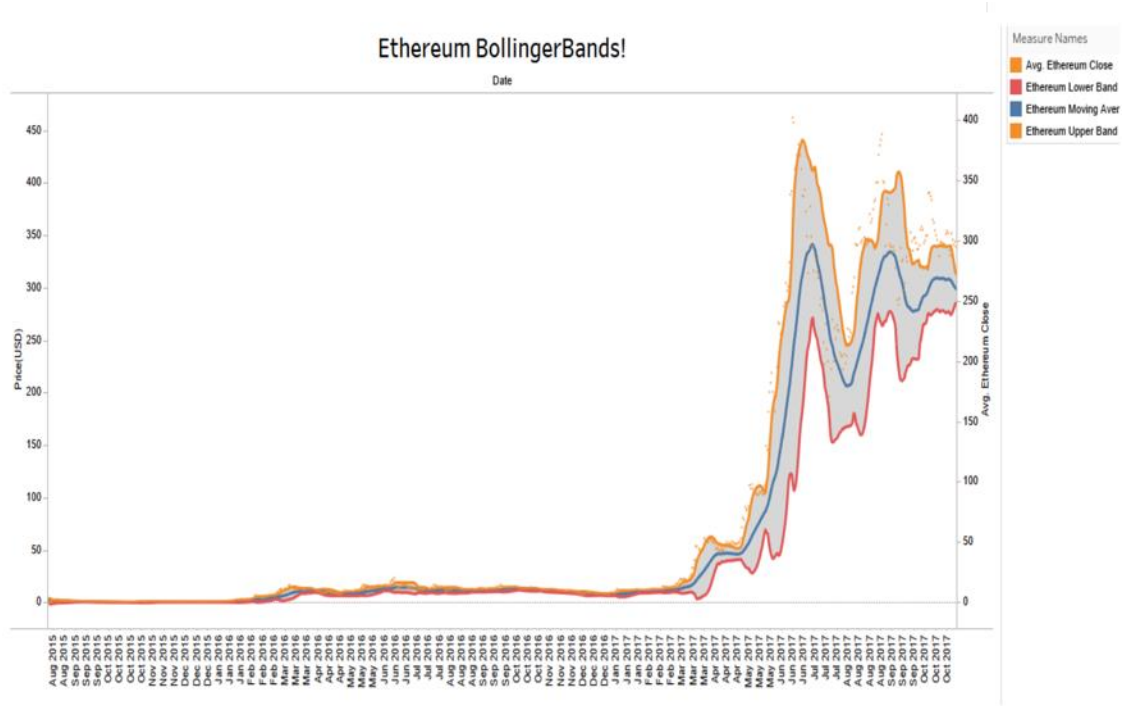
Bollinger Bands

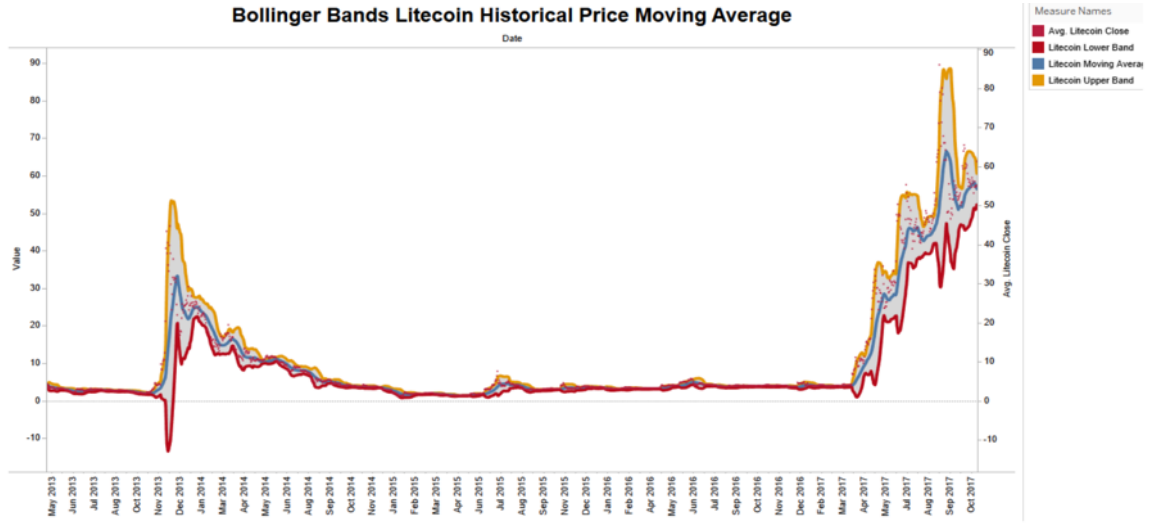
- Bollinger Bands, a popular tool among investors and traders, helps gauge the volatility of stocks and other securities to determine if they are over- or undervalued.
- Bollinger Bands is a technical analysis tool used to determine where prices are high and low
- These bands are composed of 3 lines: simple moving average (the middle band) and an upper and lower band.
- The upper and lower bands are typically two standard deviations above or below a 20-period simple moving average (SMA).
- The bands widen and narrow as the volatility of the underlying asset changes.



Bollinger bands to compare volatility







Python data visualization

- Python provides various libraries that come with different features for visualizing data. All these libraries come with different features and can support various types of graphs

- **Matplotlib**
  - Matplotlib is an easy-to-use, low-level data visualization library that is built on NumPy arrays. It consists of various plots like scatter plot, line plot, histogram, etc. Matplotlib provides a lot of flexibility.

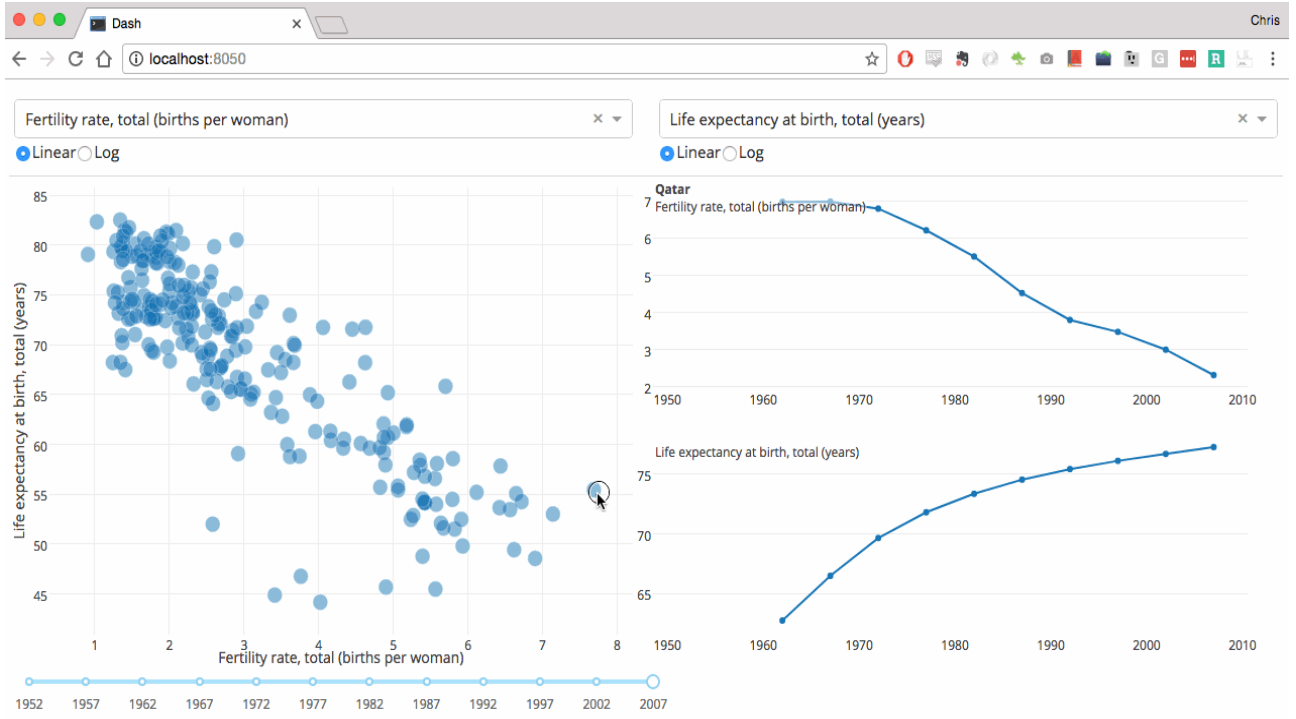
- **Seaborn**
  - Seaborn is a high-level interface built on top of the Matplotlib. It provides beautiful design styles and color palettes to make more attractive graphs.

- **Bokeh**
  - Bokeh is mainly famous for its interactive charts visualization. Bokeh renders its plots using HTML and JavaScript that uses modern web browsers for presenting elegant, concise construction of novel graphics with high-level interactivity.

- **Plotly**
  - Plotly has hover tool capabilities that allow to detect any outliers or anomalies in numerous data points.
  - It allows more customization. It makes the graph visually more attractive.

Plotly Dash

- Dash is Python framework for building web applications. It built on top of Flask, Plotly.js, React and React Js. It enables you to build dashboards using pure Python. Dash is open source, and its apps run on the web browser
- A Dash application is usually composed of two parts. The first part is the layout and describes how the app will look like and the second part describes the interactivity of the application. Dash provides HTML classes that enable us to generate HTML content with Python.
- Dash is the most downloaded, trusted Python framework for building ML & data science web apps.



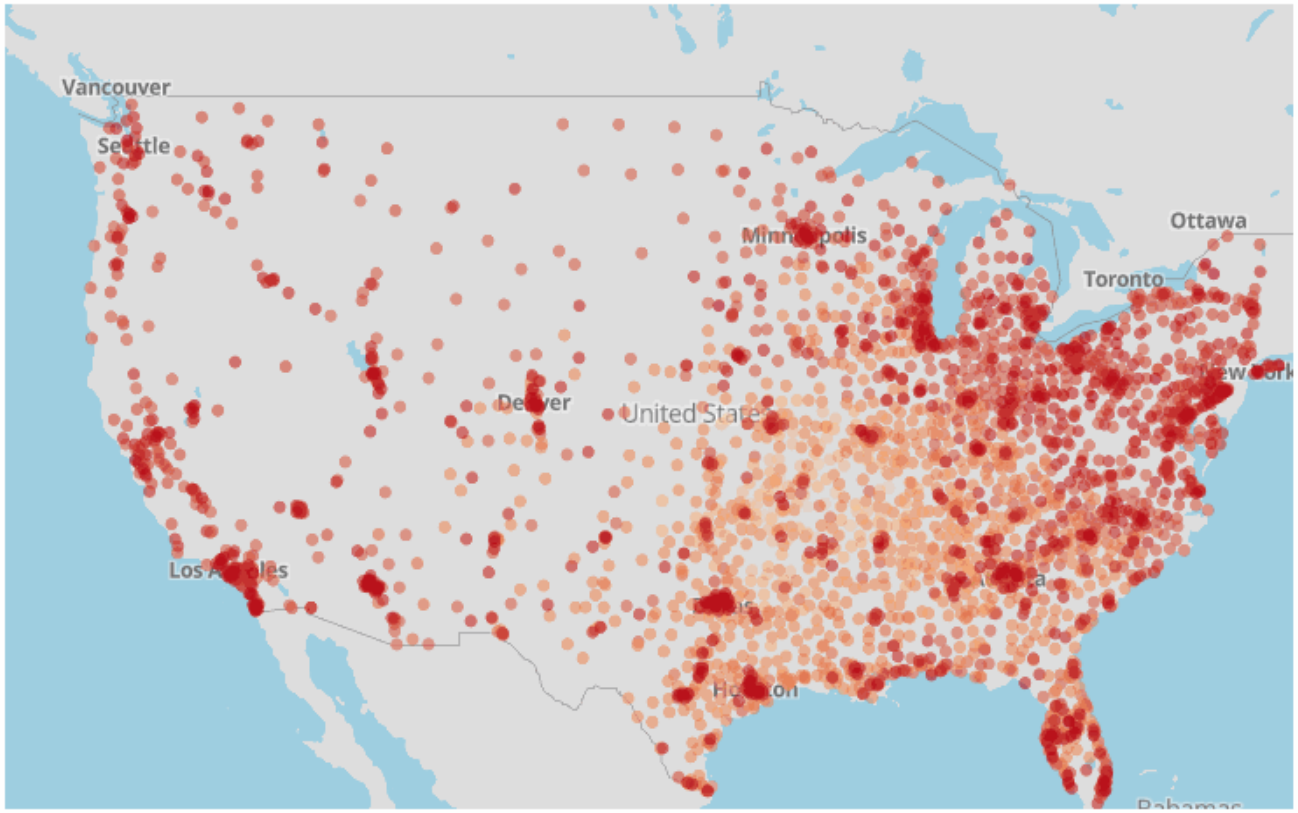
https://github.com/plotly/dash?tab=readme-ov-file

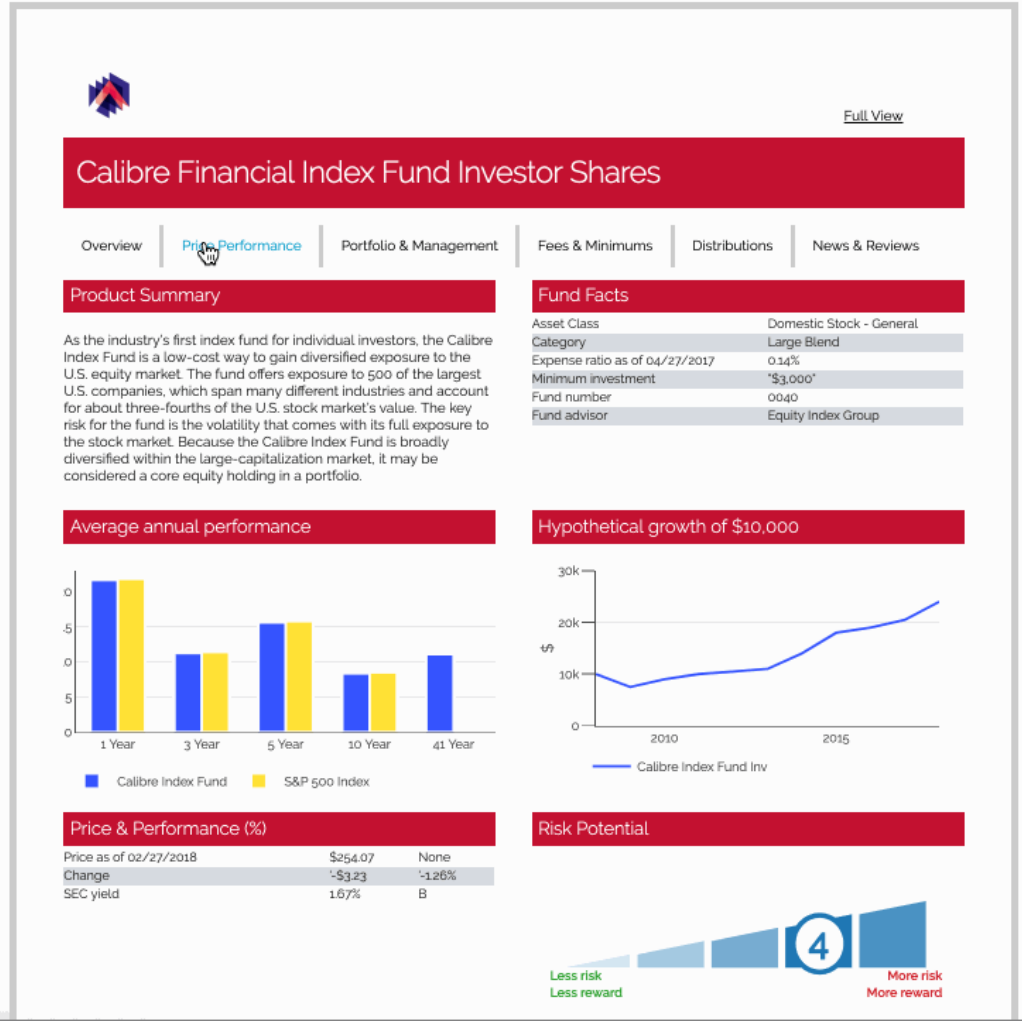
plotly | Dash

Plotly Dash

Walmart Store Openings

The Richmond, MO Supercenter opened in 1980





plotly | Dash

Plotly Dash

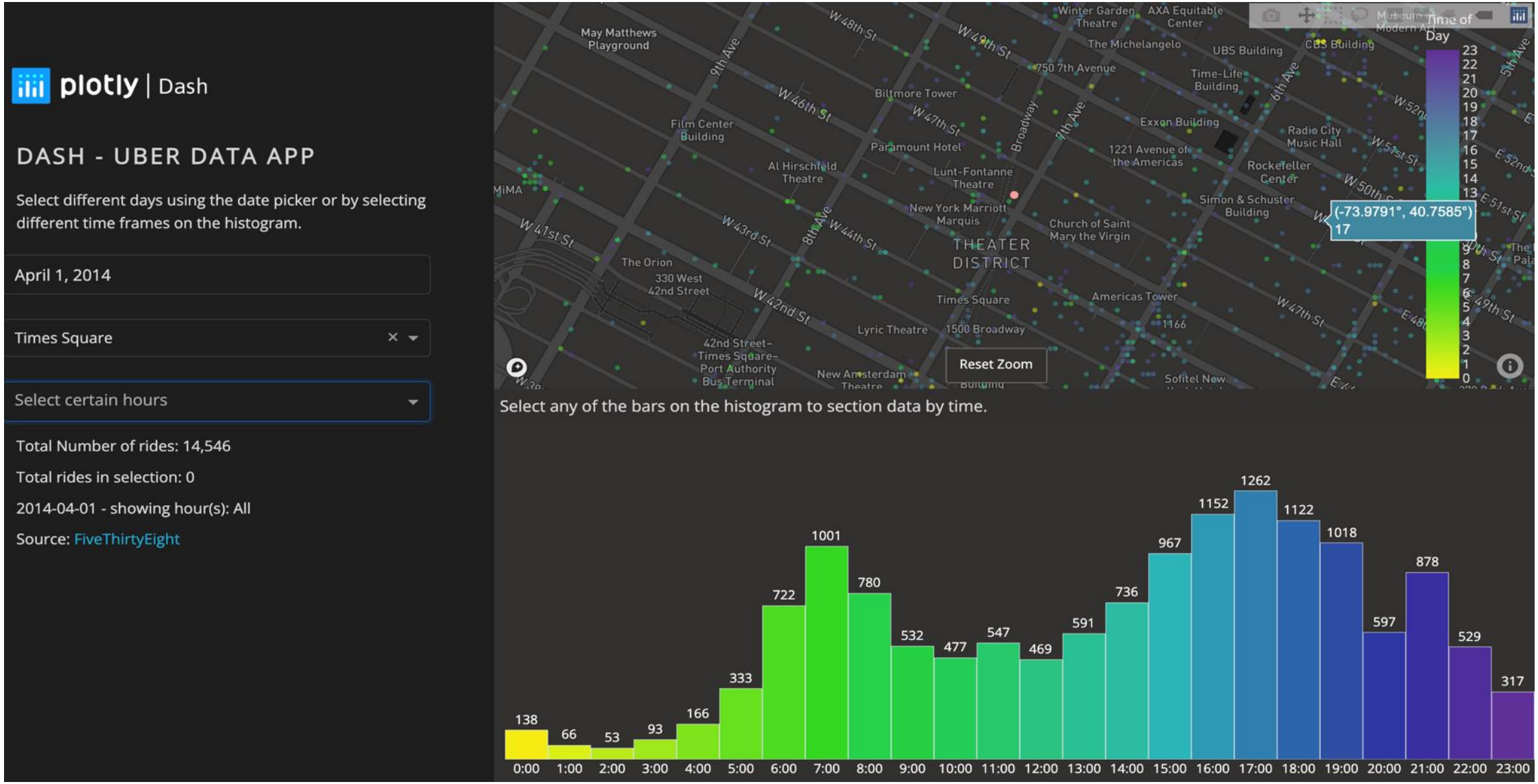
Dash Open Source + Dash Enterprise

350,000 downloads monthly

Join more than 2M Python, R, and Julia developers in building ML and data science apps.

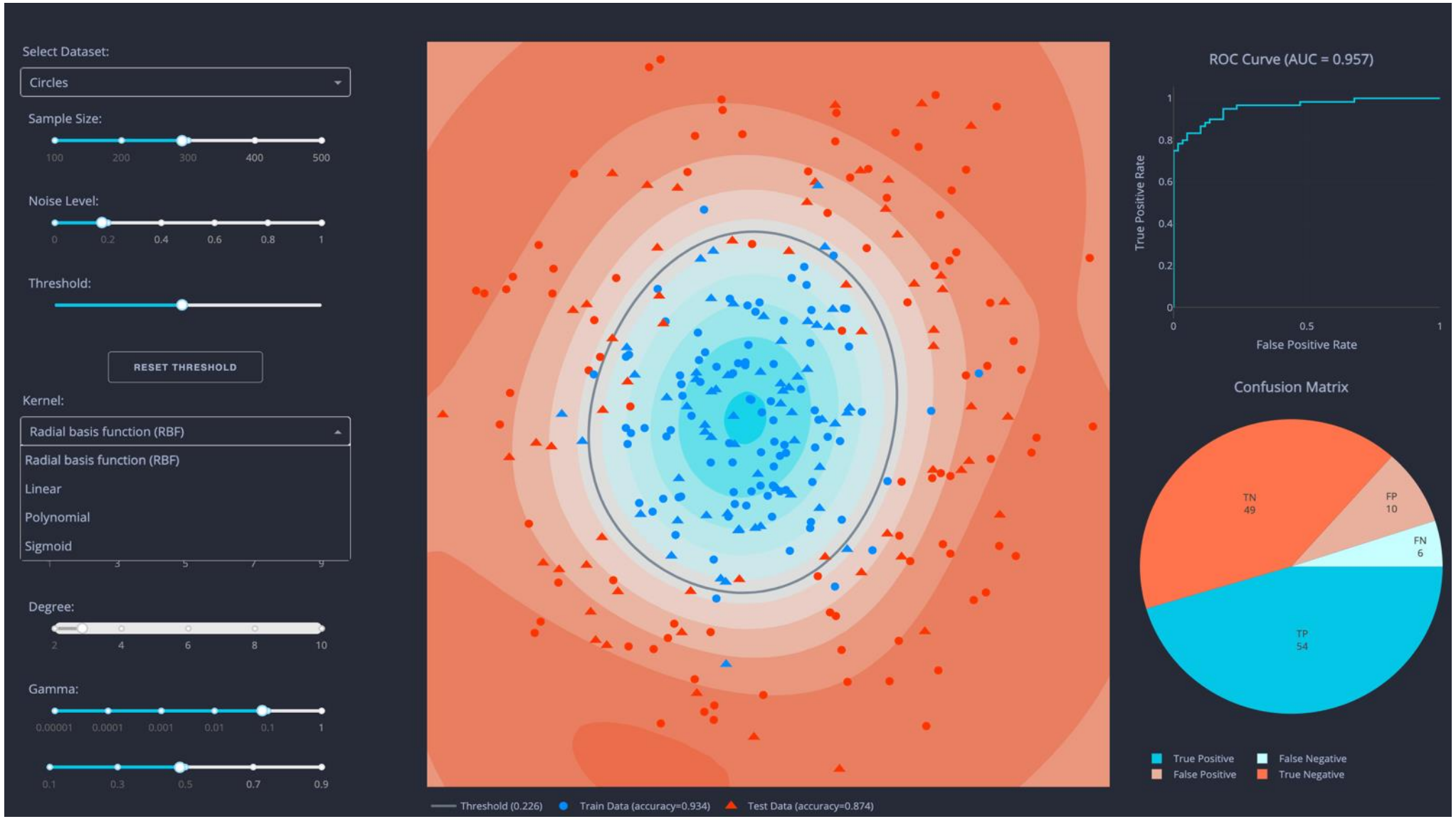
plotly | Dash

New York Uber Rides



https://dash.gallery/dash-uber-rides-demo/

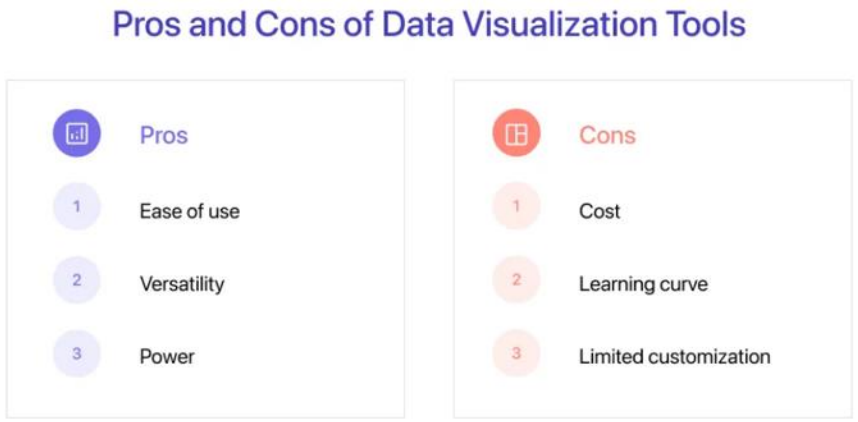
Dash is the most downloaded, trusted framework for building machine learning web apps in Python.

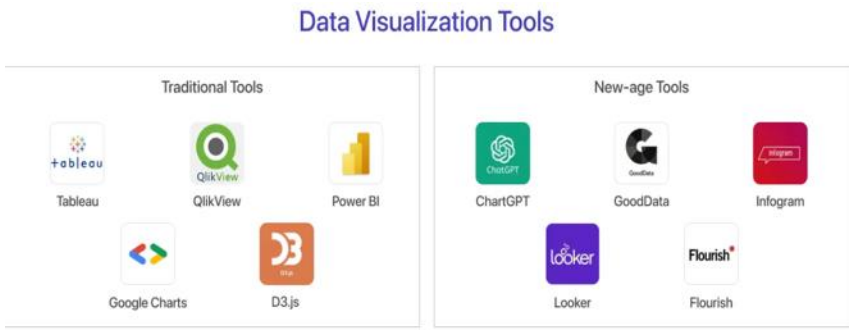


https://dash.gallery/dashr-svm/

plotly | Dash

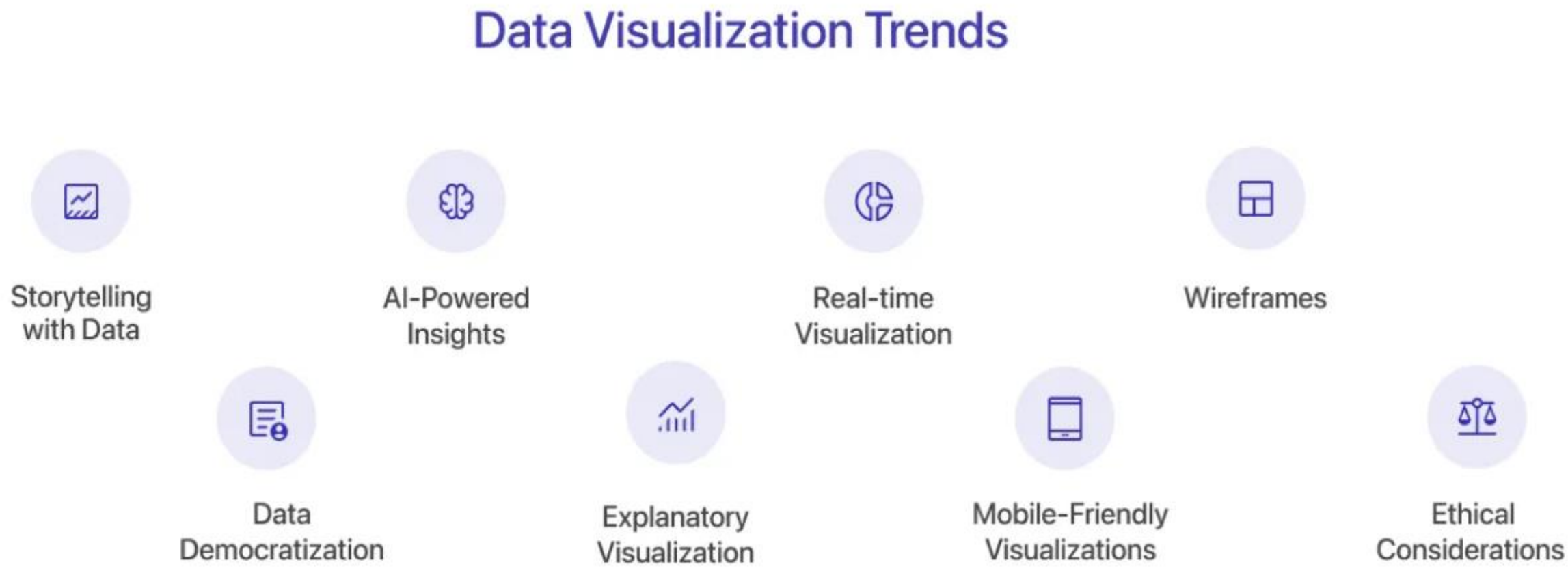
Data Visualization tools





Tool	Pros	Cons
Tableau	Easy to use, supports a wide variety of data sources, can create a wide variety of charts and graphs	Expensive
QlikView	A powerful associative engine allows users to explore data in an intuitive way	Steep learning curve
Power BI	Powerful and versatile, can be used to create a wide variety of visualizations	Large or complex datasets can lead to performance bottlenecks in Power BI
Google Charts	Free, easy to use	Limited customization options
D3.js	Powerful, can be used to create custom data visualizations	Steep learning curve

Data Visualization Trends



Conclusion

Data visualization has emerged as an indispensable tool for transforming raw data into meaningful and insightful representations. As the volume and complexity of data continue to grow exponentially, effective data visualization techniques have become essential for extracting knowledge, identifying patterns, and communicating insights to a wider audience.

Demo: Jupyter notebook demo

Data Wrangling in Python

References

- https://spectrum.adobe.com/page/data-visualization-fundamentals/
- https://www.thoughtspot.com/data-trends/data-visualization/data-visualization-principles
- https://www.effectivedatastorytelling.com/post/a-data-storytellers-guide-to-avoiding-clutter
- https://www.anychart.com/blog/2018/11/20/data-visualization-definition-history-examples/
- https://www.thoughtspot.com/data-trends/data-visualization/types-of-charts-graphs
- https://www.tableau.com/learn/articles/business-intelligence/bi-business-analytics
- https://www.dreamstime.com/seven-benefits-business-intelligence-image148961805
- https://www.scribbr.com/frequently-asked-questions/claims-supports-warrants/
- https://dash.plotly.com/layout
- https://github.com/plotly/dash?tab=readme-ov-file
- https://medium.com/@mokkup/the-future-of-data-visualization-2024-and-beyond-3173a8e60494