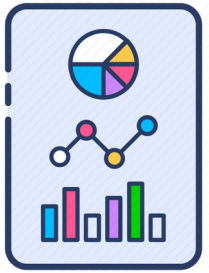
Stastical Data Visualization

Tushar B. Kute,
http://tusharkute.com



Data Visualization

- Data visualization is the presentation of data in a pictorial or graphical format.
- It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.
- With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed.

Data Visualization

- With so much information being collected through data analysis in the business world today, we must have a way to paint a picture of that data so we can interpret it.
- Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs.
- This makes the data more natural for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets.

Benefits of Data Visualization

- **Correlations in Relationships:** Without data visualization, it is challenging to identify the correlations between the relationship of independent variables. By making sense of those independent variables, we can make better business decisions.
- **Trends Over Time:** While this seems like an obvious use of data visualization, it is also one of the most valuable applications. It's impossible to make predictions without having the necessary information from the past and present. Trends over time tell us where we were and where we can potentially go.
- **Frequency:** Closely related to trends over time is frequency. By examining the rate, or how often, customers purchase and when they buy gives us a better feel for how potential new customers might act and react to different marketing and customer acquisition strategies.

Benefits of Data Visualization

- **Examining the Market:** Data visualization takes the information from different markets to give you insights into which audiences to focus your attention on and which ones to stay away from. We get a clearer picture of the opportunities within those markets by displaying this data on various charts and graphs.
- **Risk and Reward:** Looking at value and risk metrics requires expertise because, without data visualization, we must interpret complicated spreadsheets and numbers. Once information is visualized, we can then pinpoint areas that may or may not require action.
- **Reacting to the Market:** The ability to obtain information quickly and easily with data displayed clearly on a functional dashboard allows businesses to act and respond to findings swiftly and helps to avoid making mistakes.

Data Visualization Techniques

- **Infographics:** Unlike a single data visualization, infographics take an extensive collection of information and gives you a comprehensive visual representation. An infographic is excellent for exploring complex and highly-subjective topics.
- **Heatmap Visualization:** This method uses a graph with numerical data points highlighted in light or warm colors to indicate whether the data is a high-value or a low-value point. Psychologically, this data visualization method helps the viewer to identify the information because studies have shown that humans interpret colors much better than numbers and letters.

Data Visualization Techniques

- **Fever Charts:** A fever chart shows changing data over a period of time. As a marketing tool, we could take the performance from the previous year and compare that to the prior year to get an accurate projection of next year. This can help decision-makers easily interpret wide and varying data sources.
- **Area Chart (or Graph):** Area charts are excellent for visualizing the data's time-series relationship. Whether you're looking at the earnings for individual departments on a month to month basis or the popularity of a product since the 1980s, area charts can visualize this relationship.
- **Histogram:** Rather than looking at the trends over time, histograms are measuring frequencies instead. These graphs show the distribution of numerical data using an automated data visualization formula to display a range of values that can be easily interpreted.

Who uses Data Visualization?

- Data visualization is used across all industries to increase sales with existing customers and target new markets and demographics for potential customers.
- The World Advertising and Research Center (WARC) predicts that in 2020 half of the world's advertising dollars will be spent online, which means companies everywhere have discovered the importance of web data.
- As a crucial step in data analytics, data visualization gives companies critical insights into untapped information and messages that would otherwise be lost.
- The days of scouring through thousands of rows of spreadsheets are over, as now we have a visual summary of data to identify trends and patterns.

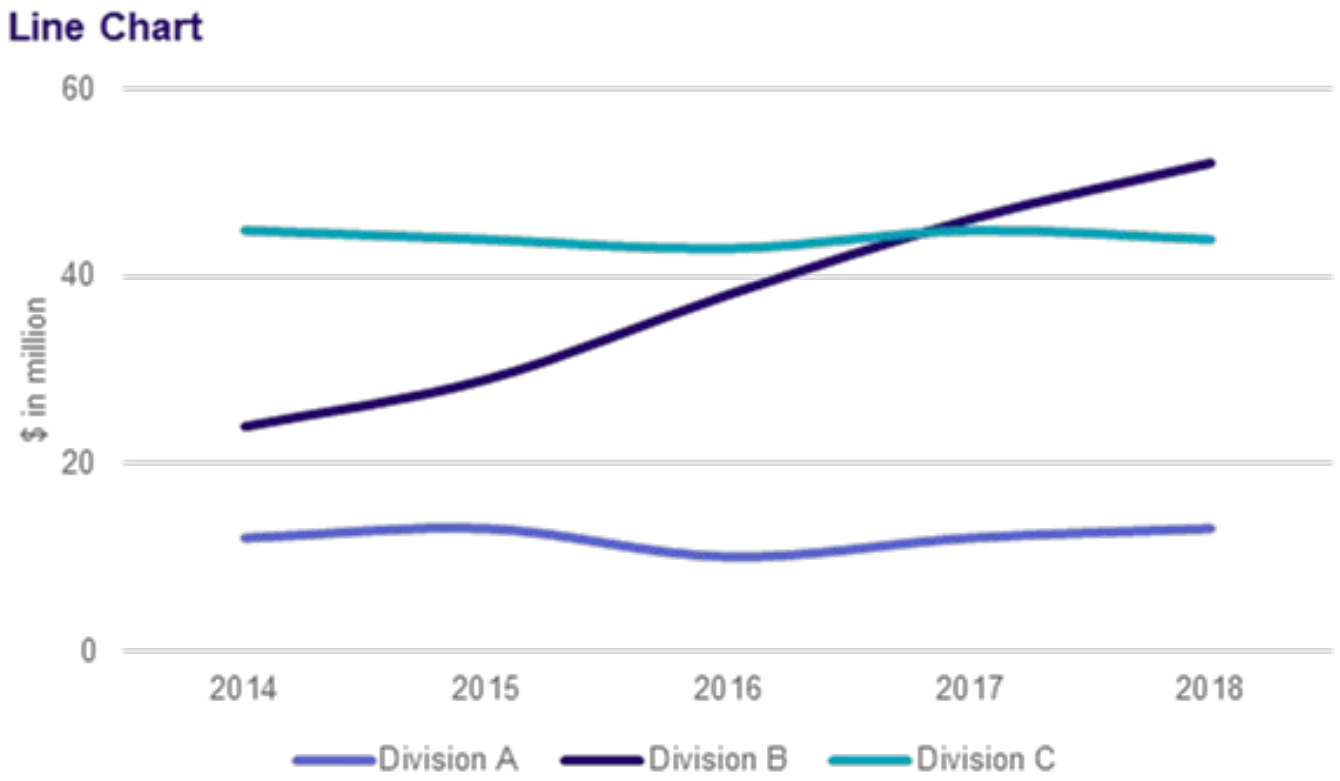
Types of Graphs

- Column Chart
- Bar Graph
- Stacked Bar Graph
- Area Chart
- Dual Axis Chart
- Line Graph
- Pie Chart
- Waterfall Chart
- Scatter Plot Chart
- Histogram
- Funnel Chart
- Heat Map

Line Chart

- A line chart is, as one can imagine, a line or multiple lines showing how single, or multiple variables develop over time. It is a great tool because we can easily highlight the magnitude of change of one or more variables over a period.
- When to use line charts
  - Remember the awkward ‘Fiction book sales’ chart we saw earlier? Well, a simple line chart would have been much better in that case.
  - A line chart allows us to track the development of several variables at the same time. It is very easy to understand, and the reader doesn’t feel overwhelmed.

Example:


Line Chart
60
40
20
0
$ in million
2014
2015
2016
2017
2018
Division A
Division B
Division C

Line Graph

- Line graphs are common and effective charts because they are simple, easy to understand, and efficient. Line charts are great for:
  - Comparing lots of data all at once
  - Showing changes and trends over time
  - Including important context and annotation
  - Displaying forecast data and uncertainty
  - Highlighting anomalies within and across data series

Line Graph

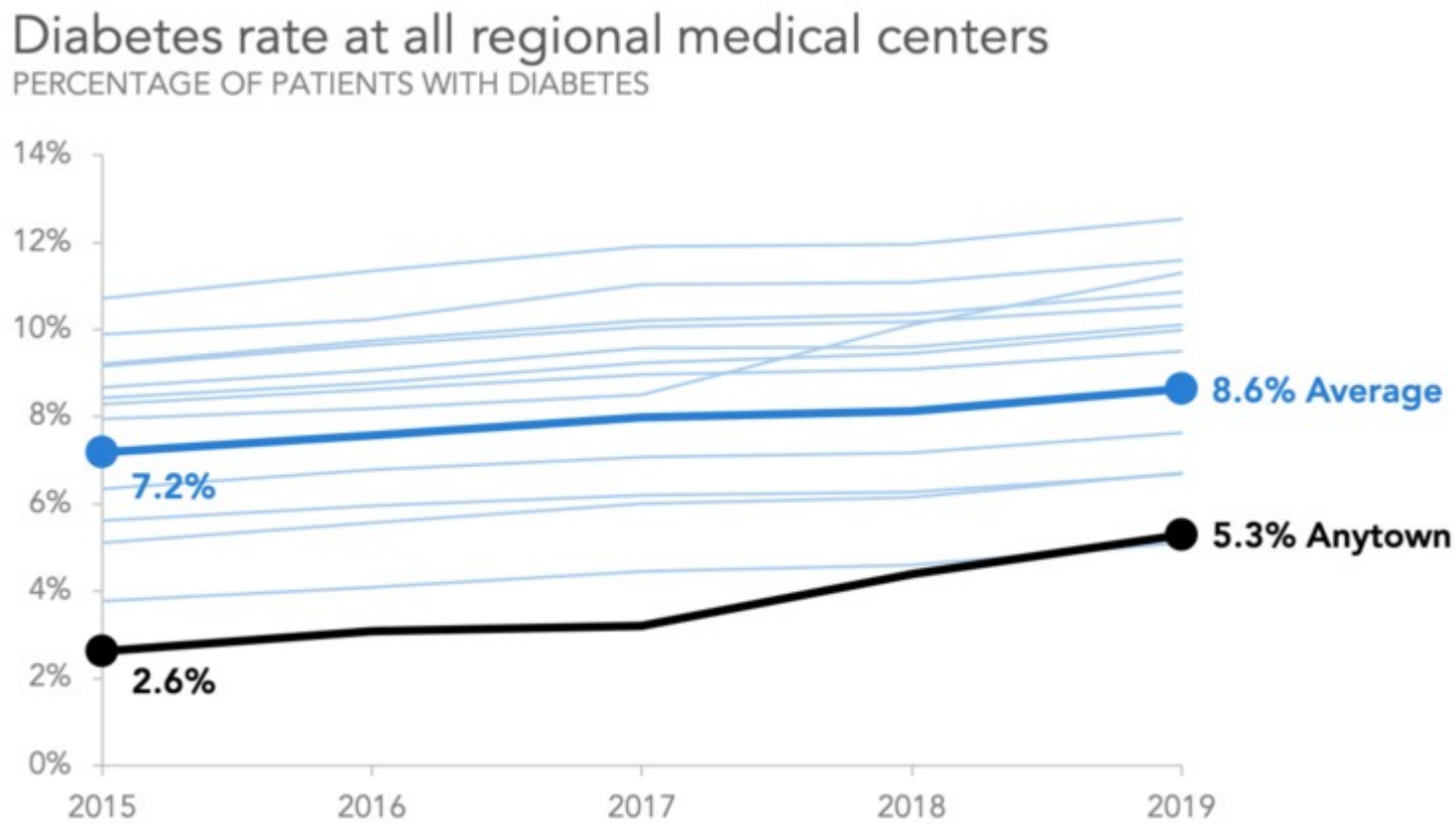
- On the other hand, they are not necessarily your best choice for:
  - Displaying quantities of things
  - Working with categorical data
  - Making part-to-whole comparisons
  - Showing sparse data sets

Line Graph


6-Month sales report and forecast
IN THOUSANDS (USD)
90
80
70
60
50
40
30
20
10
0
62.5
44.3
35.0
31.6
34.0
11.0
OCT 2019
NOV
DEC
JAN 2020
FEB
MAR

Line Graph – forecasting

Line Graph – comparison



Bar Graph

- Bar charts are among the most frequently used chart types.
- As the name suggests a bar chart is composed of a series of bars illustrating a variable’s development.
- Given that bar charts are such a common chart type, people are generally familiar with them and can understand them easily.
- Examples like this one are straightforward to read.

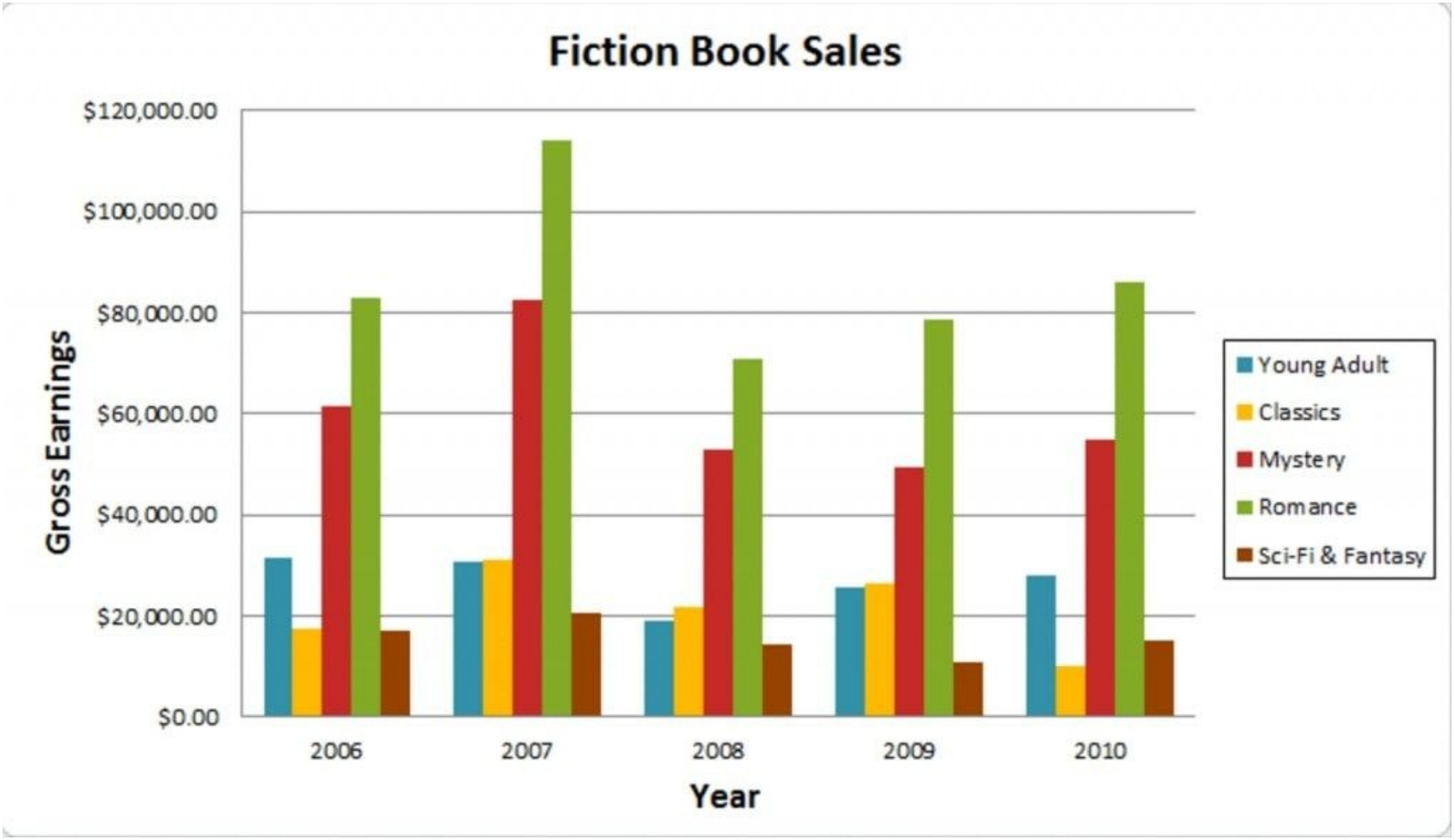
Example:


Bar Chart - One Variable
120
100
80
60
40
20
0
$ in million
81
86
91
103
109
2014
2015
2016
2017
2018
Total Revenue

Bar Graph

- However, please be aware that bar charts can be confusing, too.
- Especially if one uses them to compare several variables. I personally believe that a comparison of more than two variables with a clustered bar chart becomes too cluttered.
- Here is an example of a clustered bar chart that is not exactly crystal clear:

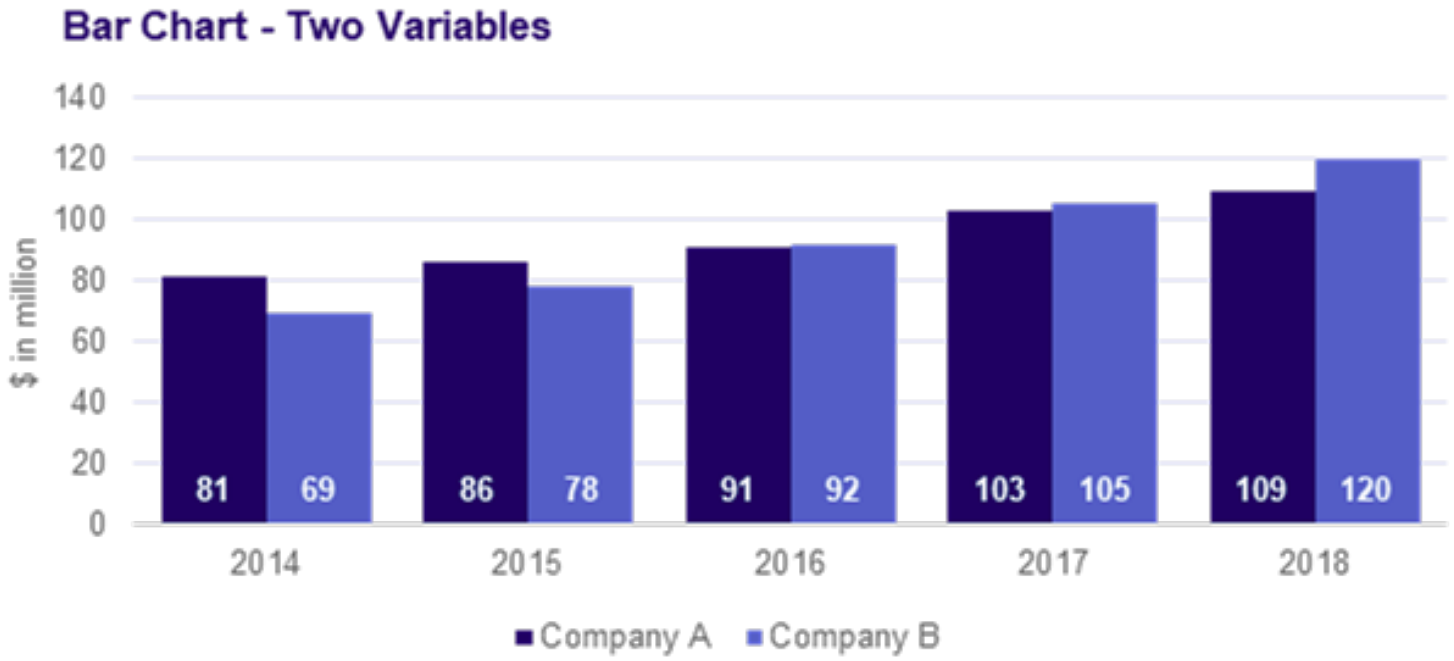
Bar Graph


Fiction Book Sales
$120,000.00
$100,000.00
$80,000.00
$60,000.00
$40,000.00
$20,000.00
$0.00
Gross Earnings
2006
2007
2008
2009
2010
Year
Young Adult
Classics
Mystery
Romance
Sci-Fi & Fantasy

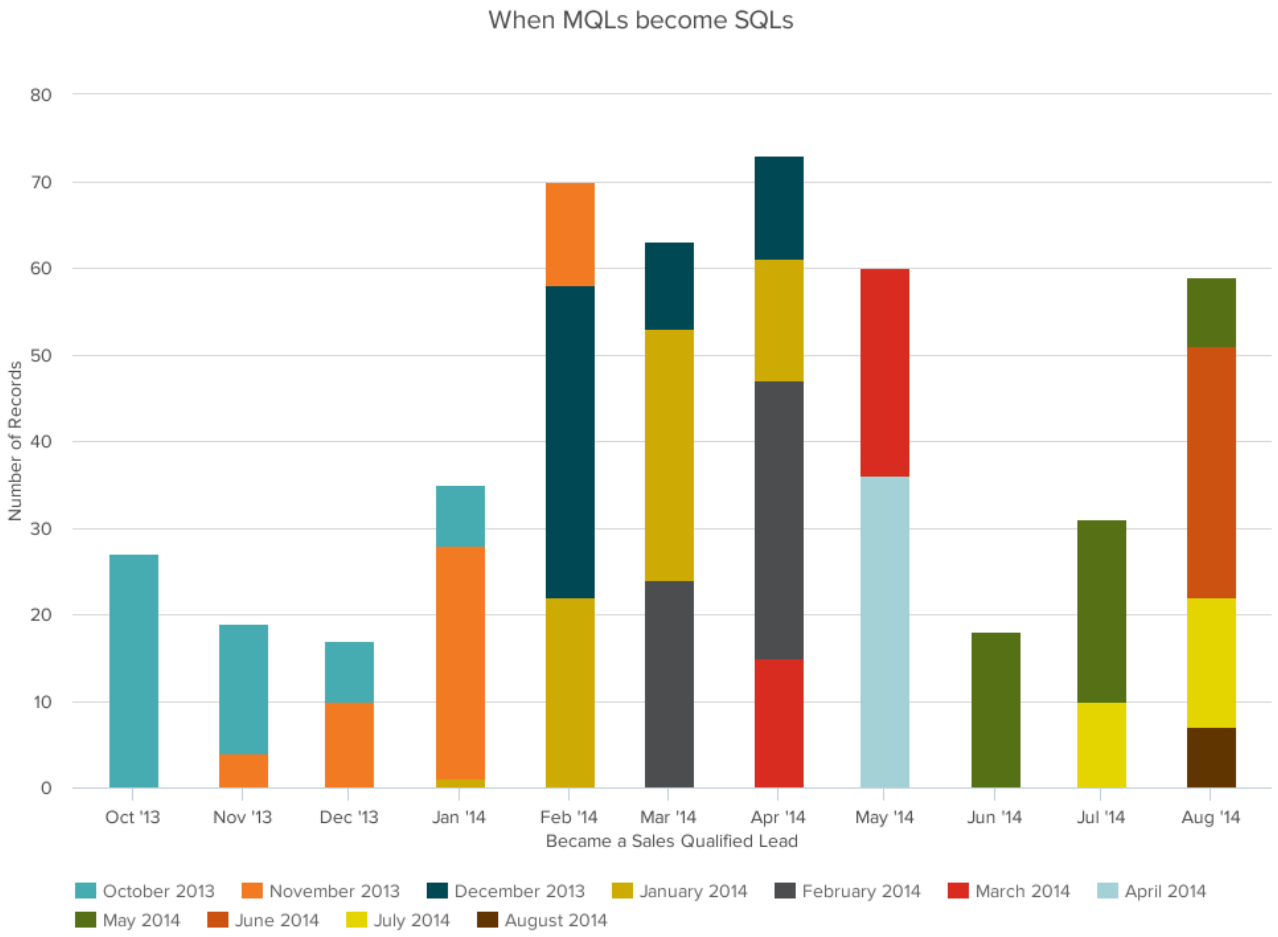
Bar Graph – When to use ?

- Bar charts are nice but limited. We have to consider the type of data we want to visualize and the number of variables that will be added to the chart.
- Bar charts are great when we want to track the development of one or two variables over time.
- For example, one of the most frequent applications of bar charts in corporate presentations is to show how a company's total revenues have developed during a given period.

Bar Chart – two variables



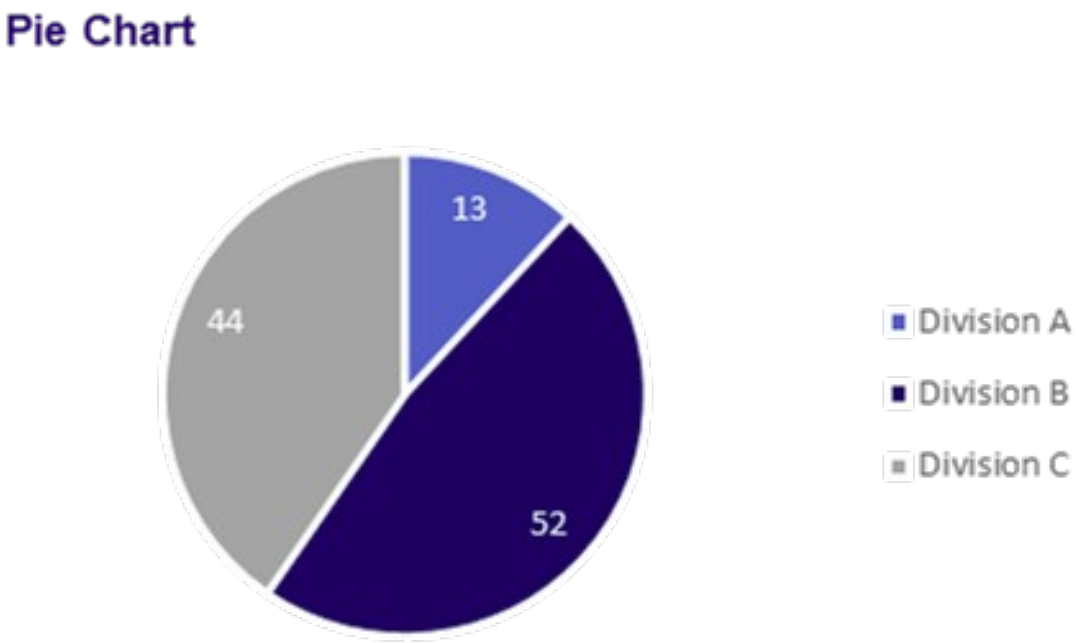
Stacked Bar Chart



Pie Chart

- A pie chart is a circular graph divided into slices. The larger a slice is the bigger portion of the total quantity it represents.
- When to use a pie chart?
  - So, pie charts are best suited to depict sections of a whole.
- What does that mean?
  - If a company operates three separate divisions, at year-end its top management would be interested in seeing what portion of total revenue each division accounted for.

Example:



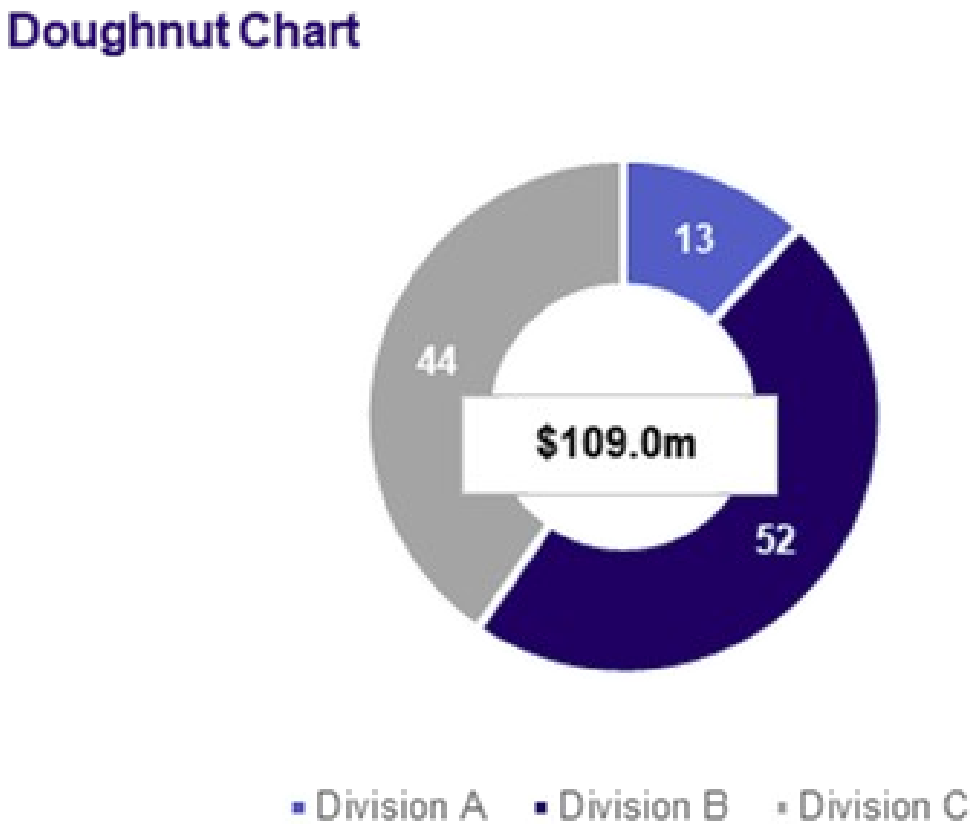
Pie Chart – When to avoid?

- Obviously, we can't use a pie chart in situations when we would like to show how one or more variables develop over time.
- Pie charts are a definite no-go in these cases. Moreover, as mentioned earlier, a pie chart would be misleading if we don't consider all values.
- In the context of our example from earlier, we shouldn't create a pie chart that includes revenue of only two of the firm's three divisions.

Doughnut Chart

- Doughnut charts are basically pie charts with a hole in the middle. (It is as if their heart is missing…)
- When to use doughnut charts
  - The use cases of pie and doughnut charts are identical.
  - The only important difference is that doughnut charts allow us to indicate the total amount by adding a text box in the middle.

Example:


Doughnut Chart
13
44
$109.0m
52
Division A
Division B
Division C

Doughnut Chart – When to avoid?

- We already explained when to avoid using pie charts. The same is valid for doughnut charts.
- One piece of advice when choosing whether to include a pie or a doughnut chart would be to think of your audience. How likely is it they would be interested in seeing the total figure for breakdown you are providing? If the split itself is more important, then go ahead and use a pie chart. If the value of the total is important too, then perhaps a doughnut chart would be preferable.
- Moreover, some studies have shown that people tend to get a distorted idea when shown pie charts, as larger portions can look even more so because they cover more space. With doughnut charts, this isn't as much of an issue.

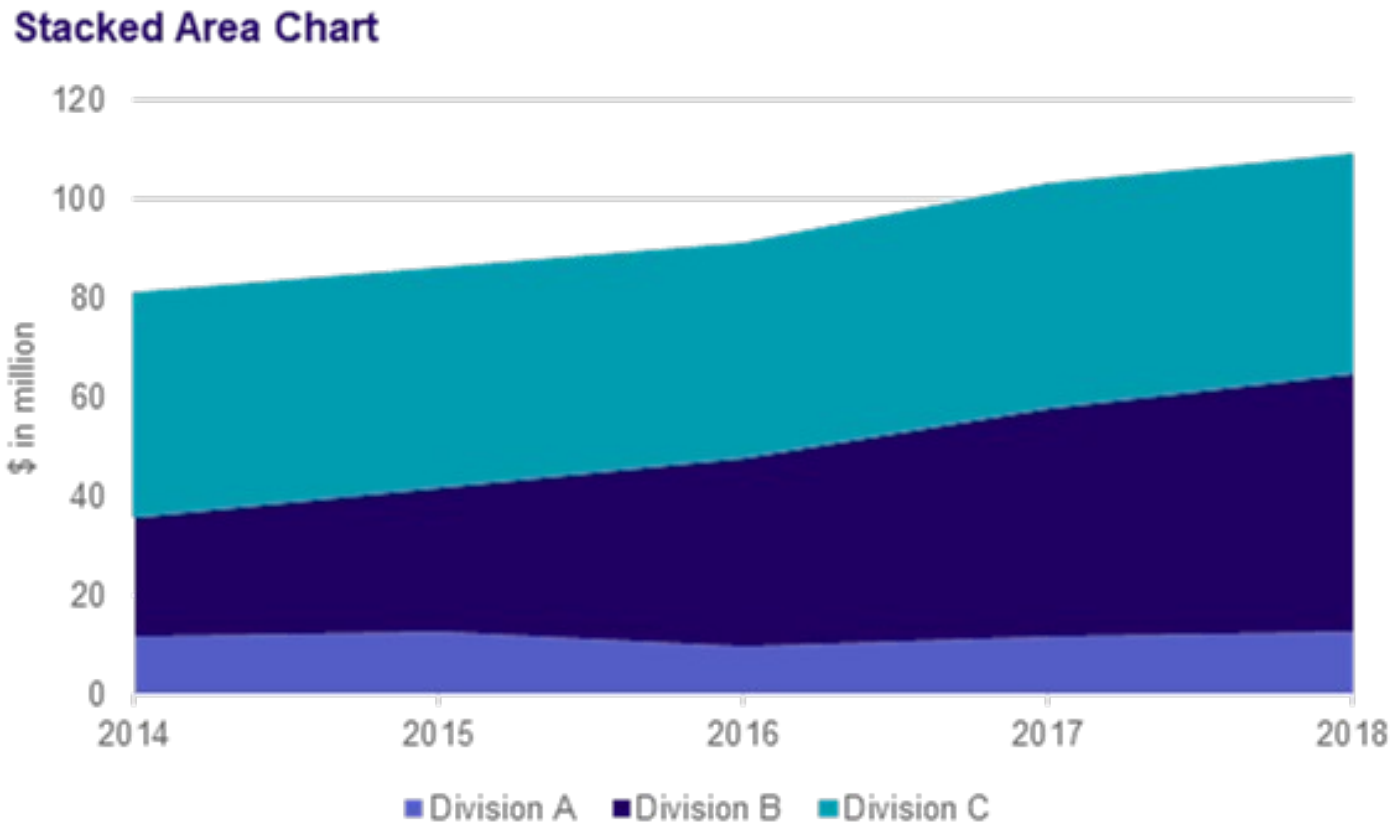
Area Chart

- Area charts are very similar to line charts.
- The idea of an area chart is based on the line chart. Coloured regions (areas) show us the development of each variable over time.
- There are three types of area charts: regular area chart, stacked area chart, and 100% stacked area chart.

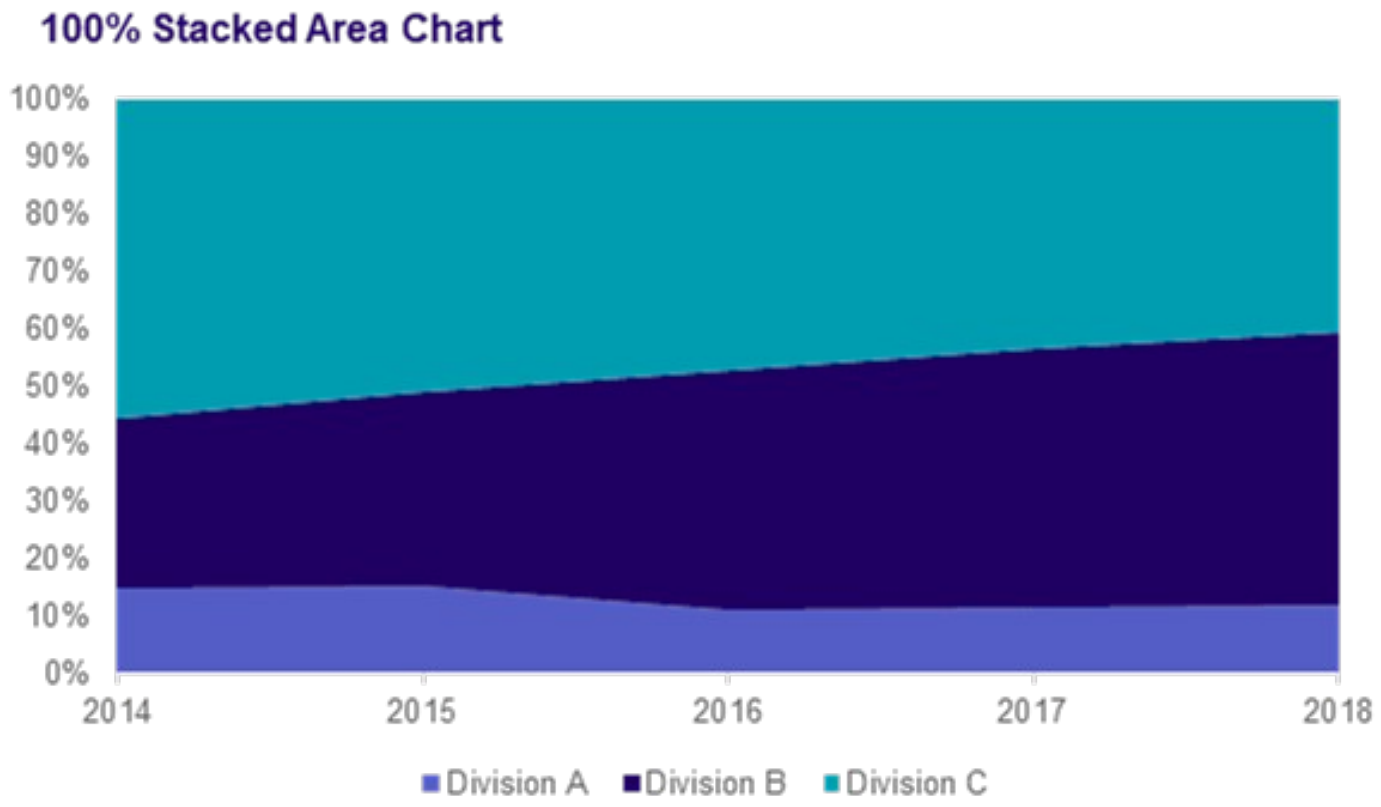
When to use Area Chart?

- Whenever we want to show how the parts of a whole change over time, we should consider an area chart. So, for example, if the company has three revenue generating divisions, it is very likely that management would like to see the development of each of these divisions.
- This is a great way to draw attention to the total value and still emphasize an important trend – say, revenues from one division have been growing rapidly while the other two have kept the same level. A stacked area chart is perfect in this case.
- However, if we are interested in the portion of revenue generated by each division and not that much of the total amount of revenues, we can simply use a 100% stacked area chart. This will show each division's percentage contribution over time.

Example:


Stacked Area Chart
120
100
80
60
40
20
0
$ in million
2014
2015
2016
2017
2018
Division A
Division B
Division C

Example:

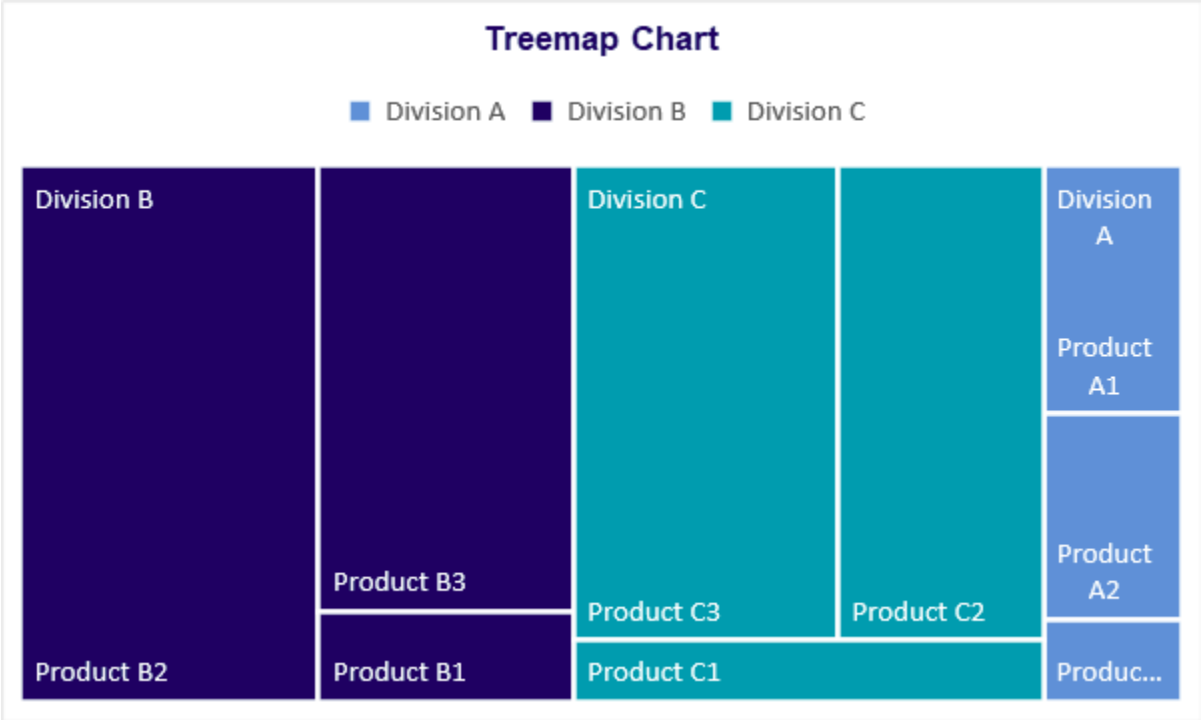

100% Stacked Area Chart
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2014
2015
2016
2017
2018
Division A
Division B
Division C

Area Chart – When to avoid?

- Obviously, similarly to line charts, area charts are not suitable for representing parts of a whole over a single period.
- In our example, we can't use an area chart to show the proportion of revenues each division generated in say, 2018 alone. So that's a situation where we can't use an area chart.
- In general, I would stay away from the classical area chart too. It can be very confusing and even Microsoft themselves recommend avoiding it and to consider using a simple line chart.

Treemap Chart

- There are some chart types that are effective but often neglected. Treemap charts are a good example. Here is what one looks like.



Treemap – when to use?

- The company we have been looking at so far has three divisions. And each of them has its own products.
- This is the perfect way to provide information about the weight divisions have with respect to the firm's total revenue.
- At the same time, it shows how much each product contributes to the revenue of its division.

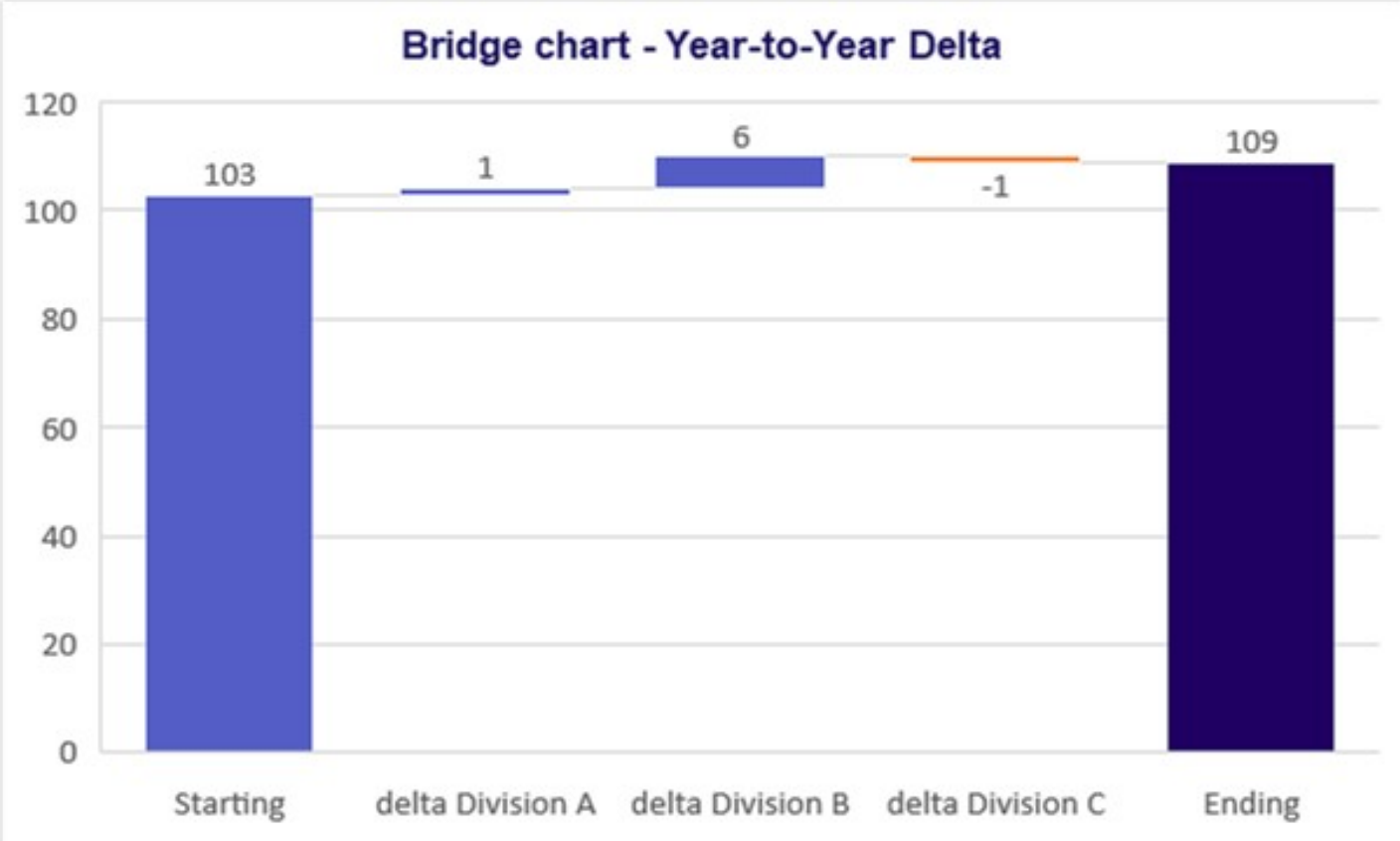
Treemap – when to avoid?

- As you can imagine it is quite difficult to apply treemap charts to a context that is not the one we just described.
- Treemap charts are one of the chart types that are not suitable when the data we are working with is not divisible into categories and sub-categories.
- Moreover, we can’t use treemap charts if we want to track development over time.

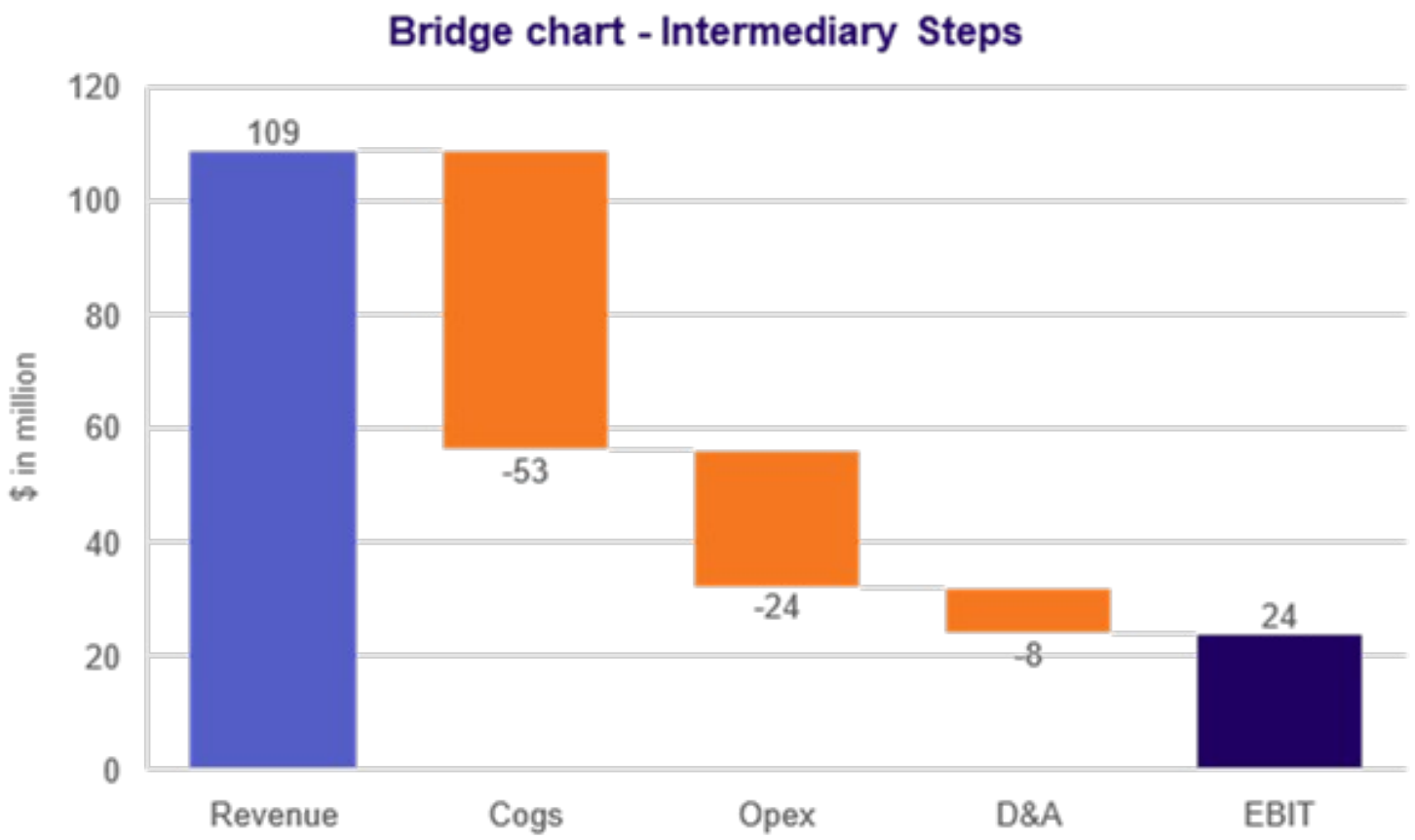
Waterfall Chart

- Waterfall, also known as bridge charts, take their origins from consulting.
- Several decades ago top tier “24/7 at your service” consultants at McKinsey popularized this type of visualization among their clients. And ever since, the popularity of bridge charts has continued to rise.
- Bridge charts are made of bars showing the cumulative effect of a series of positive and negative values impacting a starting and an ending value.

Example:


Bridge chart - Year-to-Year Delta
120
100
80
60
40
20
0
103
1
6
-1
109
Starting
delta Division A
delta Division B
delta Division C
Ending

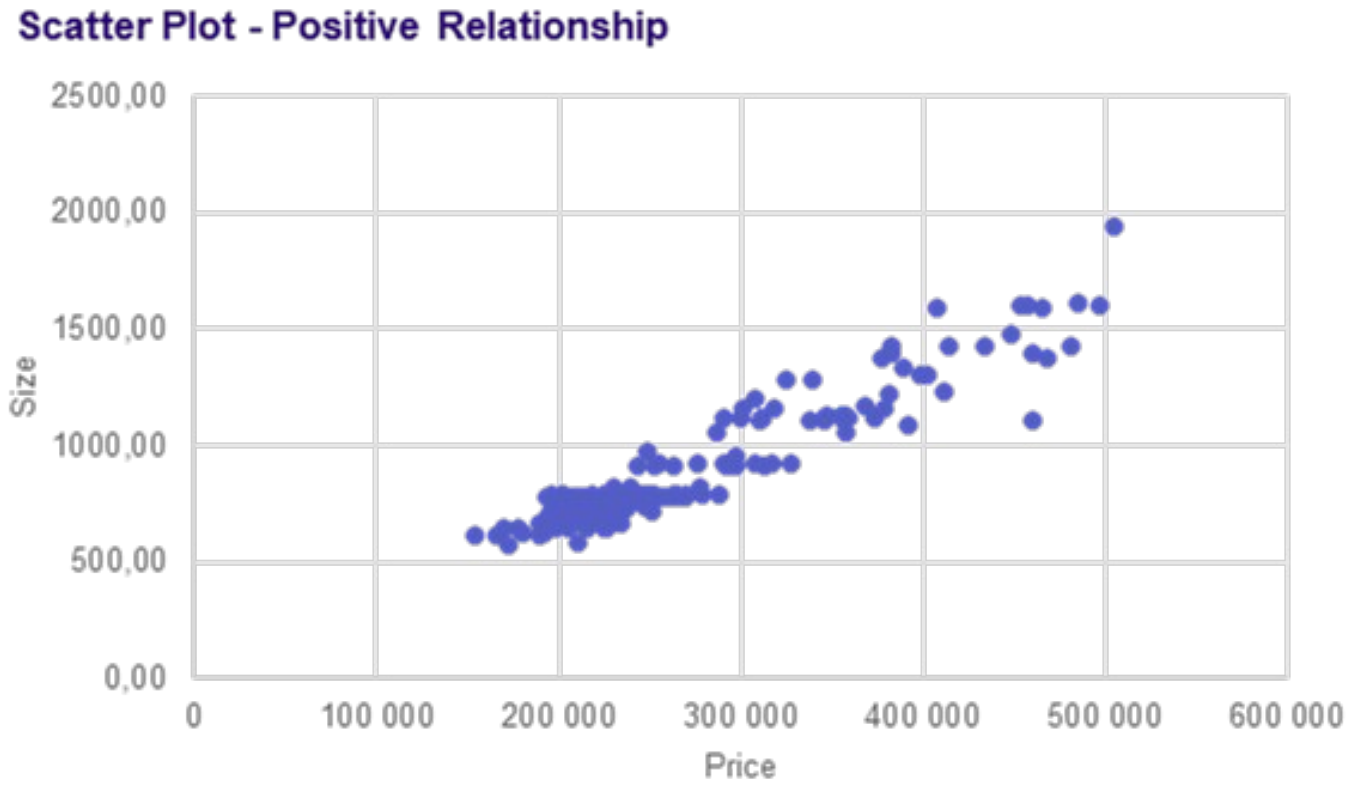
Example:


Bridge chart - Intermediary Steps
$ in million
120
100
80
60
40
20
0
109
-53
-24
-8
24
Revenue
Cogs
Opex
D&A
EBIT

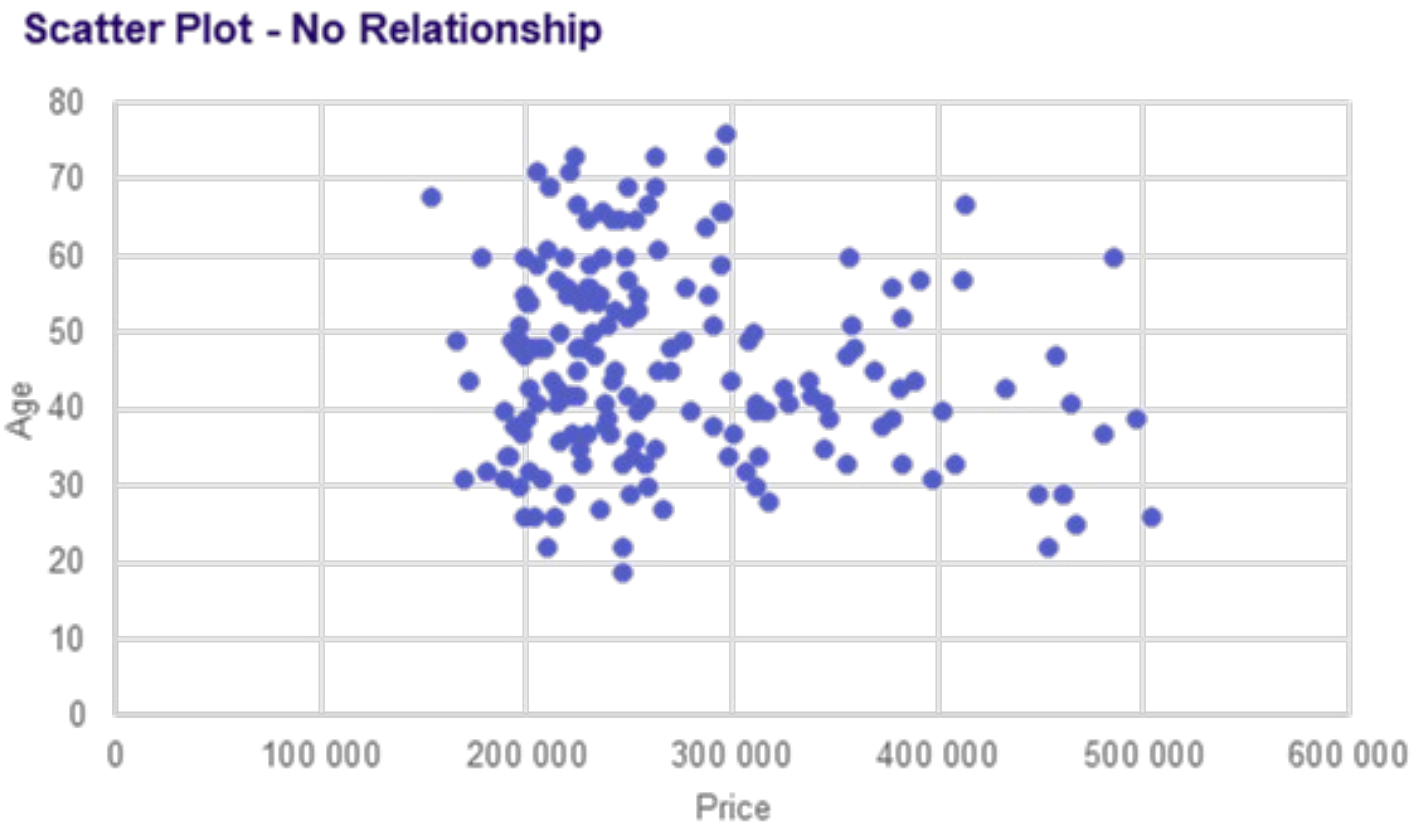
Scatter Plot

- A scatter plot is a type of chart that is often used in the fields of statistics and data science. It consists of multiple data points plotted across two axes.
- Each variable depicted in a scatter plot would have multiple observations. If a scatter plot includes more than two variables, then we would use different colours to signify that.
- When to use scatter plots
  - A scatter plot chart is a great indicator that allows us to see whether there is a pattern to be found between two variables.

Example:


Scatter Plot - Positive Relationship
2500,00
2000,00
1500,00
1000,00
500,00
0,00
Size
0
100 000
200 000
300 000
400 000
500 000
600 000
Price

Example:


Scatter Plot - No Relationship
Age
80
70
60
50
40
30
20
10
0
0
100 000
200 000
300 000
400 000
500 000
600 000
Price

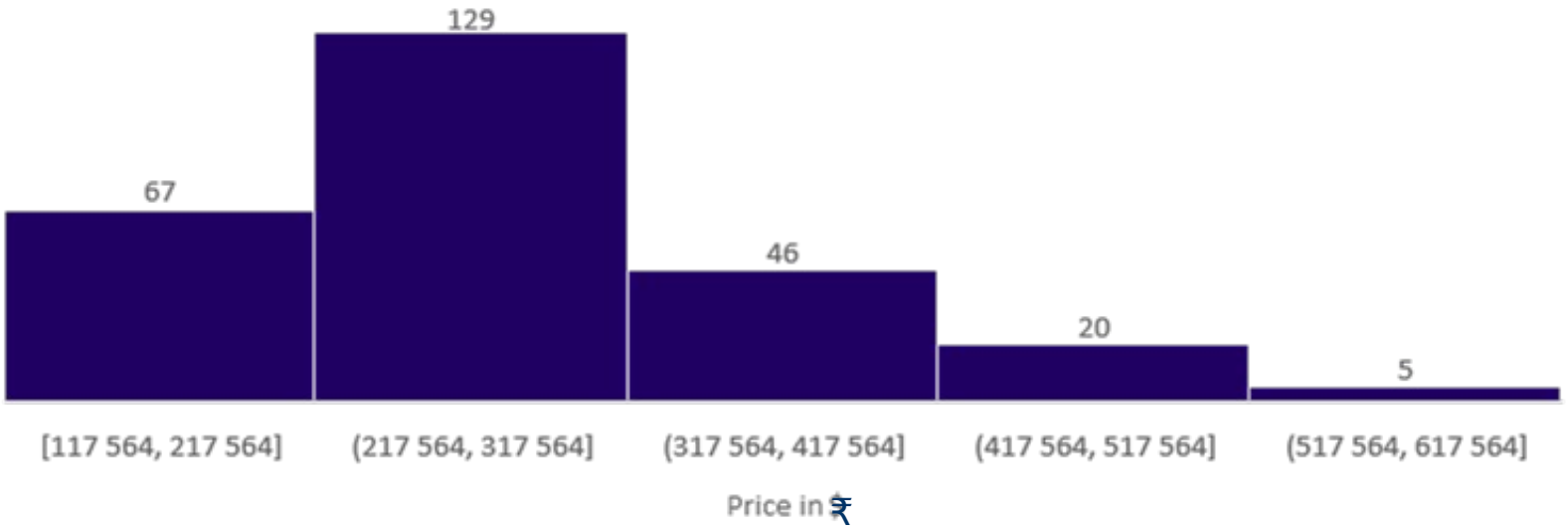
Scatter Plot – When to avoid?

- We can’t use scatter plots when we don’t have bi-dimensional data.
- In our example, we need information about both house prices and house size to create a scatter plot. A scatter plot requires at least two dimensions for our data.
- In addition, scatter plots are not suitable if we are interested in observing time patterns.
- Finally, a scatter plot is used with numerical data, or numbers. If we have categories such as 3 divisions, 5 products, and so on, a scatter plot would not reveal much.

Histogram

- A series of bins showing us the frequency of observations of a given variable. The definition of histogram charts is short and easy.
- Here's an example.
  - An interviewer asked 267 people how much their house cost. Then a histogram was used to portray the interviewer's findings. Some prices were in the range between ₹117-217k, many more in the range ₹217-₹317k, and the rest of the houses were classified in more expensive bins.

Example:

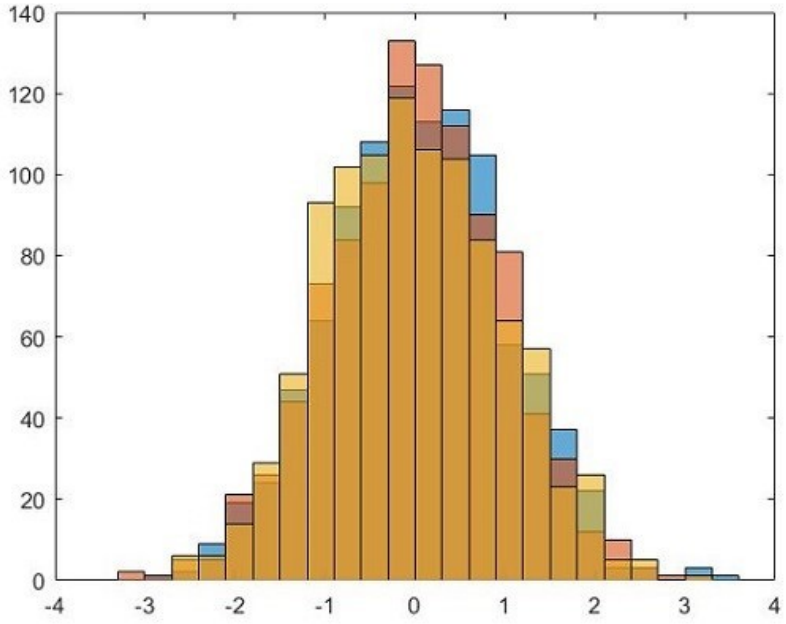

129
67
46
20
5
[117 564, 217 564]
(217 564, 317 564]
(317 564, 417 564]
(417 564, 517 564]
(517 564, 617 564]
Price in ₹

Histogram – When to use?

- Histograms are great when we would like to show the distribution of the data we are working with.
- This allows us to group continuous data into bins and hence, provide a useful representation of where observations are concentrated.

Histogram – When to avoid?

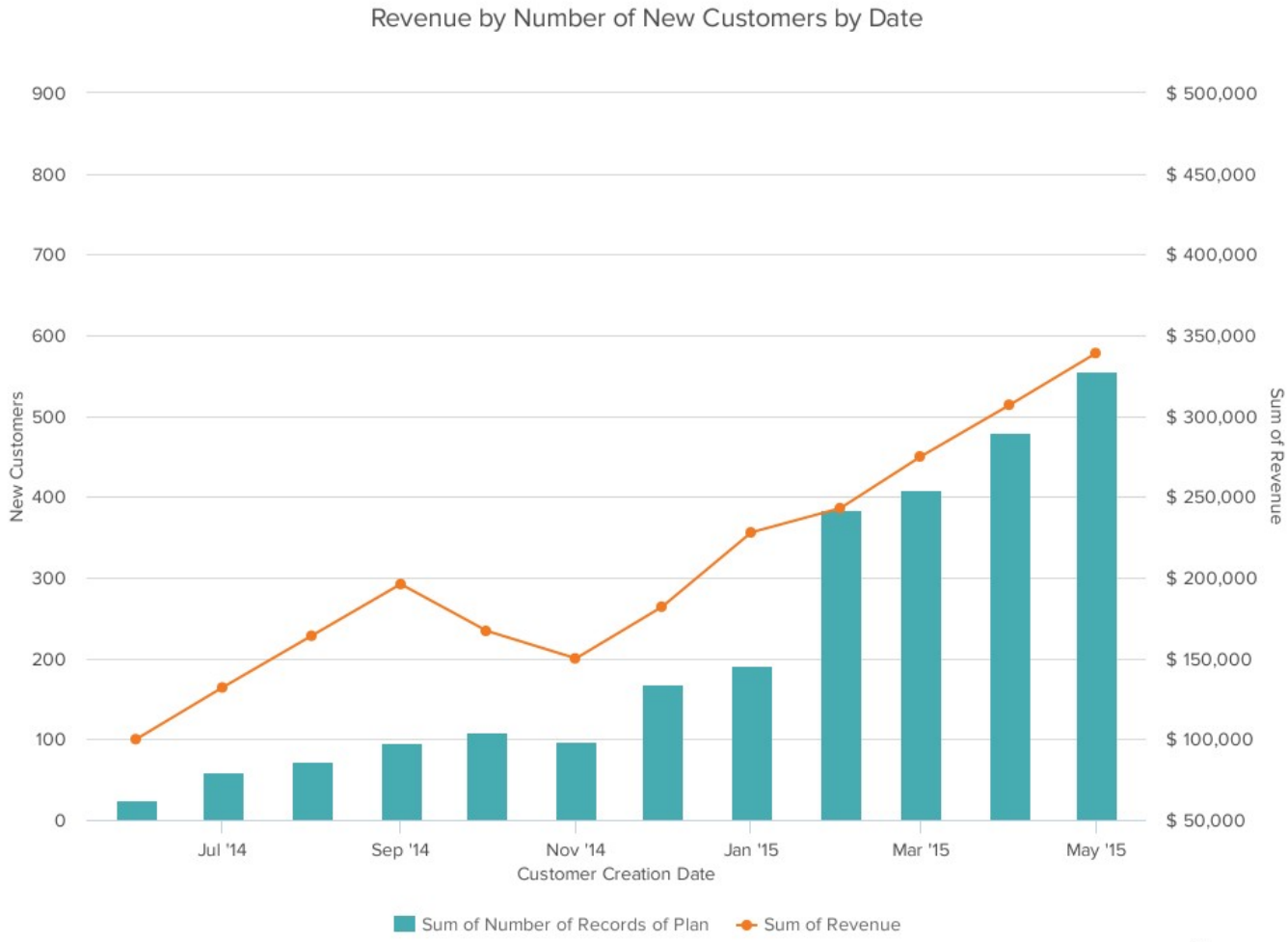
- Be careful when the data you are working with contains multiple categories or variables. Multi-column histograms are among the chart types to be avoided when they look like this.



Dual Axis Chart

- A dual axis chart allows you to plot data using two y-axes and a shared x-axis.
- It's used with three data sets, one of which is based on a continuous set of data and another which is better suited to being grouped by category.
- This should be used to visualize a correlation or the lack thereof between these three data sets.

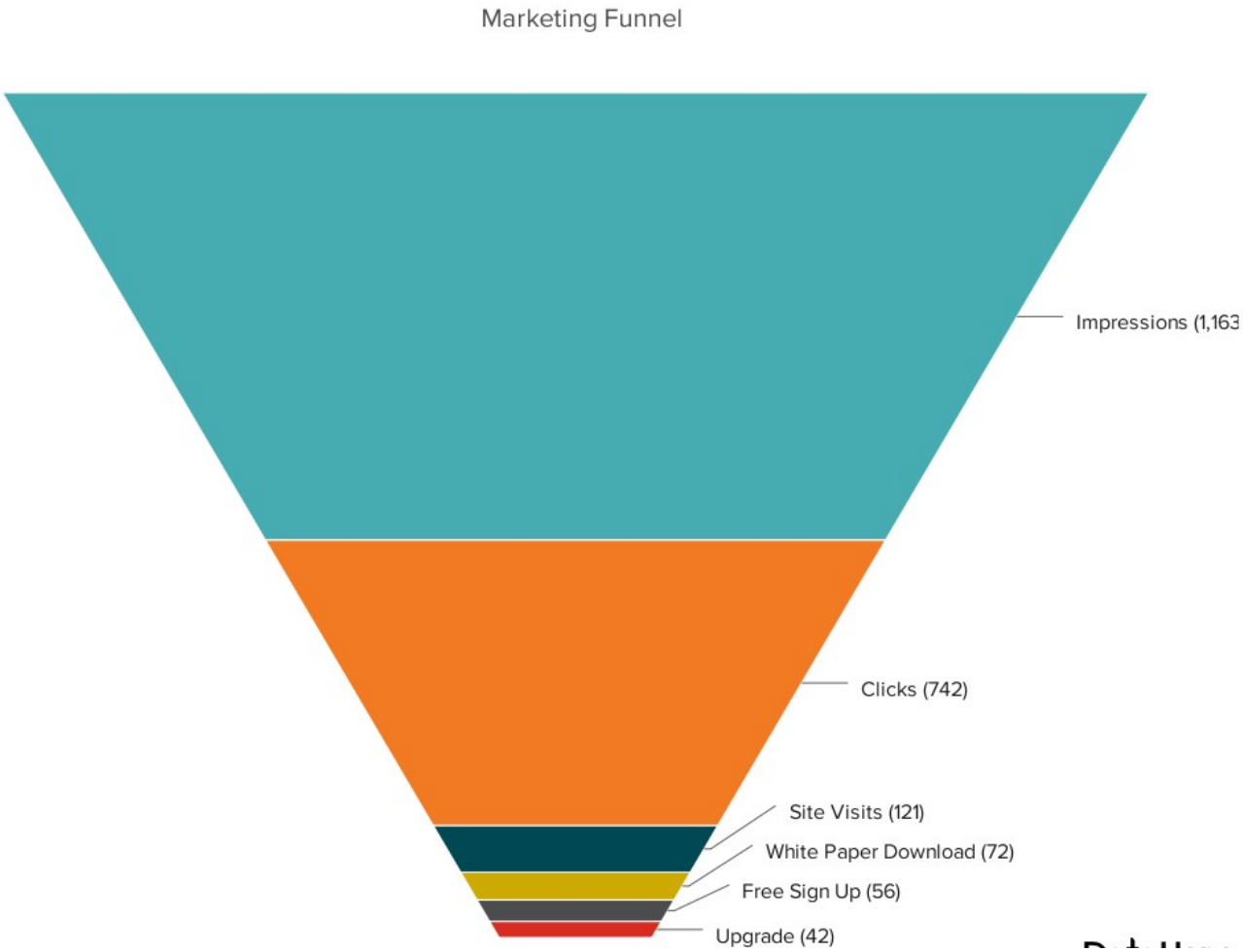
Example:


Revenue by Number of New Customers by Date
900
800
700
600
500
400
300
200
100
0
New Customers
$ 500,000
$ 450,000
$ 400,000
$ 350,000
$ 300,000
$ 250,000
$ 200,000
$ 150,000
$ 100,000
$ 50,000
Sum of Revenue
Jul '14
Sep '14
Nov '14
Jan '15
Mar '15
May '15
Customer Creation Date
Sum of Number of Records of Plan
Sum of Revenue

Funnel Chart

- A funnel chart shows a series of steps and the completion rate for each step.
- This can be used to track the sales process or the conversion rate across a series of pages or steps.

Example:


Marketing Funnel
Impressions (1,163
Clicks (742)
Site Visits (121)
White Paper Download (72)
Free Sign Up (56)
Upgrade (42)

Heatmap

- A heat map shows the relationship between two items and provides rating information, such as high to low or poor to excellent.
- The rating information is displayed using varying colors or saturation.

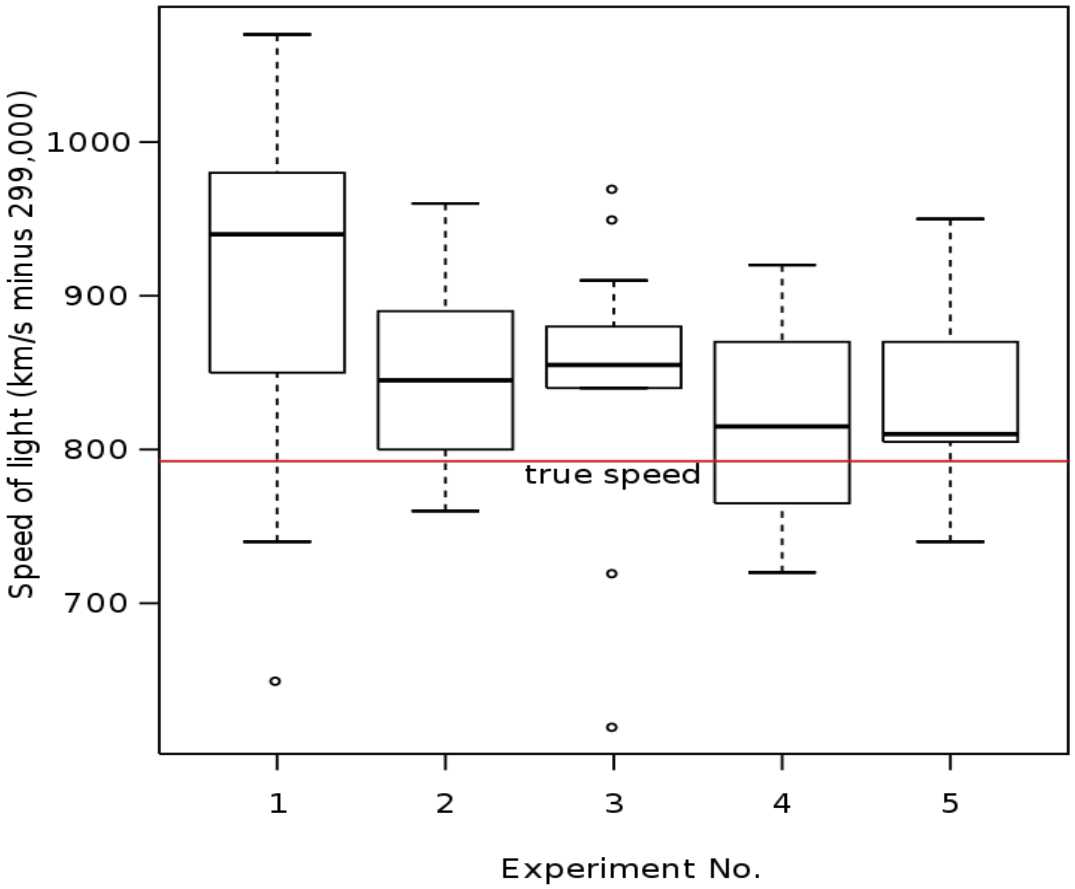
Example:


Highest Degree vs. Class Identification
HIGHEST DEGREE
0 - Less than HS
1 - High School
2 - Junior College
3 - Bachelor
4 - Graduate
67
144
78
8
91
552
337
17
9
87
72
4
6
90
230
23
2
32
136
21
1 - lower class
2 - working class
3 - middle class
4 - upper class
SUBJECTIVE CLASS IDENTIFICATION

Box Plot

- A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles.
- Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.
- Outliers may be plotted as individual points.

Box Plot


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

Data Visualization Tools

- Tableau
- Infogram
- ChartBlocks
- D3.js
- Google Charts
- Fusion Charts
- Chart.js

Tableau

- Tableau has a variety of options available, including a desktop app, server and hosted online versions, and a free public option.
- There are hundreds of data import options available, from CSV files to Google Ads and Analytics data to Salesforce data.
- Output options include multiple chart formats as well as mapping capability. That means designers can create color-coded maps that showcase geographically important data in a format that's much easier to digest than a table or chart could ever be.

Infogram

- Infogram is a fully-featured drag-and-drop visualization tool that allows even non-designers to create effective visualizations of data for marketing reports, infographics, social media posts, maps, dashboards, and more.
- Finished visualizations can be exported into a number of formats: .PNG, .JPG, .GIF, .PDF, and .HTML. Interactive visualizations are also possible, perfect for embedding into websites or apps.
- Infogram also offers a WordPress plugin that makes embedding visualizations even easier for WordPress users.

ChartBlocks

- ChartBlocks claims that data can be imported from “anywhere” using their API, including from live feeds. While they say that importing data from any source can be done in “just a few clicks,” it’s bound to be more complex than other apps that have automated modules or extensions for specific data sources.
- The app allows for extensive customization of the final visualization created, and the chart building wizard helps users pick exactly the right data for their charts before importing the data.
- Designers can create virtually any kind of chart, and the output is responsive—a big advantage for data visualization designers who want to embed charts into websites that are likely to be viewed on a variety of devices.

D3.js

- D3.js is a JavaScript library for manipulating documents using data.
- D3.js requires at least some JS knowledge, though there are apps out there that allow non-programming users to utilize the library.
- Those apps include NVD3, which offers reusable charts for D3.js; Plotly's Chart Studio, which also allows designers to create WebGL and other charts; and Ember Charts, which also uses the Ember.js framework.

Google Charts

- Google Charts is a powerful, free data visualization tool that is specifically for creating interactive charts for embedding online.
- It works with dynamic data and the outputs are based purely on HTML5 and SVG, so they work in browsers without the use of additional plugins. Data sources include Google Spreadsheets, Google Fusion Tables, Salesforce, and other SQL databases.
- There are a variety of chart types, including maps, scatter charts, column and bar charts, histograms, area charts, pie charts, treemaps, timelines, gauges, and many others. These charts can be customized completely, via simple CSS editing.

FusionCharts

- FusionCharts is another JavaScript-based option for creating web and mobile dashboards. It includes over 150 chart types and 1,000 map types.
- It can integrate with popular JS frameworks (including React, jQuery, React, Ember, and Angular) as well as with server-side programming languages (including PHP, Java, Django, and Ruby on Rails).
- FusionCharts gives ready-to-use code for all of the chart and map variations, making it easier to embed in websites even for those designers with limited programming knowledge.

Chart.js

- Chart.js is a simple but flexible JavaScript charting library. It’s open source, provides a good variety of chart types (eight total), and allows for animation and interaction.
- Chart.js uses HTML5 Canvas for output, so it renders charts well across all modern browsers. Charts created are also responsive, so it’s great for creating visualizations that are mobile-friendly.

Visualization using Programming

- Python
  - matplotlib
  - seaborn
  - plotly
  - pylab
- R
  - graphics
  - ggplot2

Best Resources to Learn

- https://python-graph-gallery.com
- https://www.r-graph-gallery.com

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com