STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering

PES UNIVERSITY ONLINE

STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation - Scatterplot

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Scatter plot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables.

The position of each dot on the horizontal and vertical axis indicates values for an individual data point.

The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Source: https://chartio.com/learn/charts/what-is-a-scatter-plot/

When to use Scatter plot?

Scatter plots are used to observe relationships between variables (numeric).

Identification of correlational relationships are common with scatter plots.

Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points.

Source: https://chartio.com/learn/charts/what-is-a-scatter-plot/

Scatter Plot - Explained

The items in the population may have several values associated with them.

For example, imagine choosing a random sample of days and determining the average temperature and humidity on each day.

Each day in the population provides two values, temperature and humidity. The random sample therefore would consist of pairs of numbers.

If the precipitation were measured on each day as well, the sample would consist of triplets.

In principle, any number of quantities could be measured on each day, producing a sample in which each item is a list of numbers.

Text Book – Chapter 1.3 – Pg. No - 37

Multivariate and Bivariate Data

Data for which each item consists of more than one value is called **multivariate data.**

When each item is a pair of values, the data are said to be bivariate.

One of the most useful graphical summaries for numerical bivariate data is the **scatterplot.**

Text Book – Chapter 1.3 – Pg. No - 37

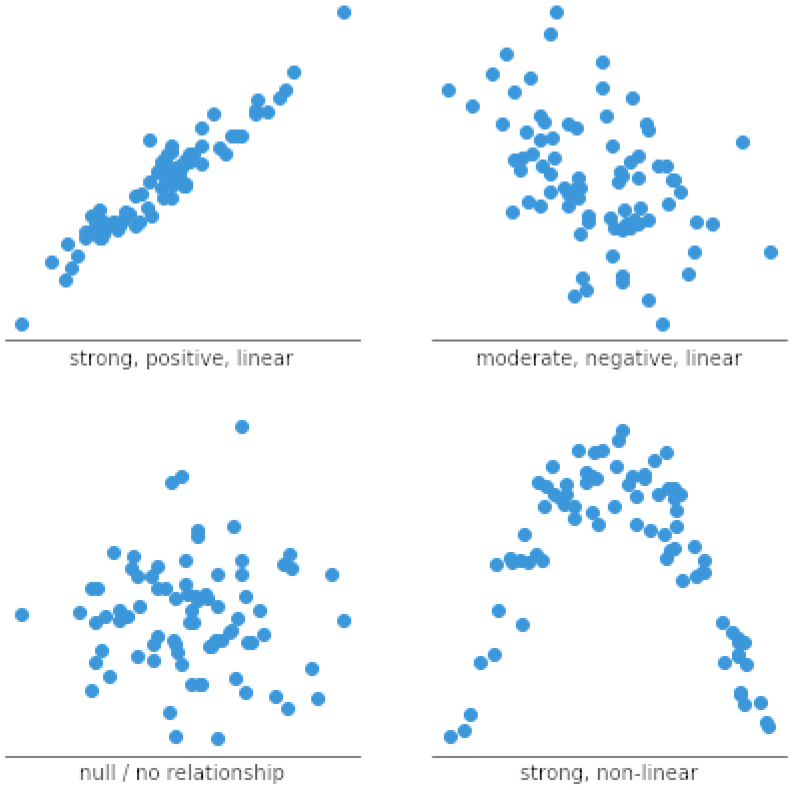
How to construct Scatter Plot?

- Collect the two paired sets of data.
- Create a summary table of the data.
- Draw and label the horizontal and vertical axes.
- Plot the data pairs on the diagram by placing a dot at the
- intersection of each data pair.
- Look at how the two variables vary together.

Source: www.slideshare.net

Relationships in Scatter Plot

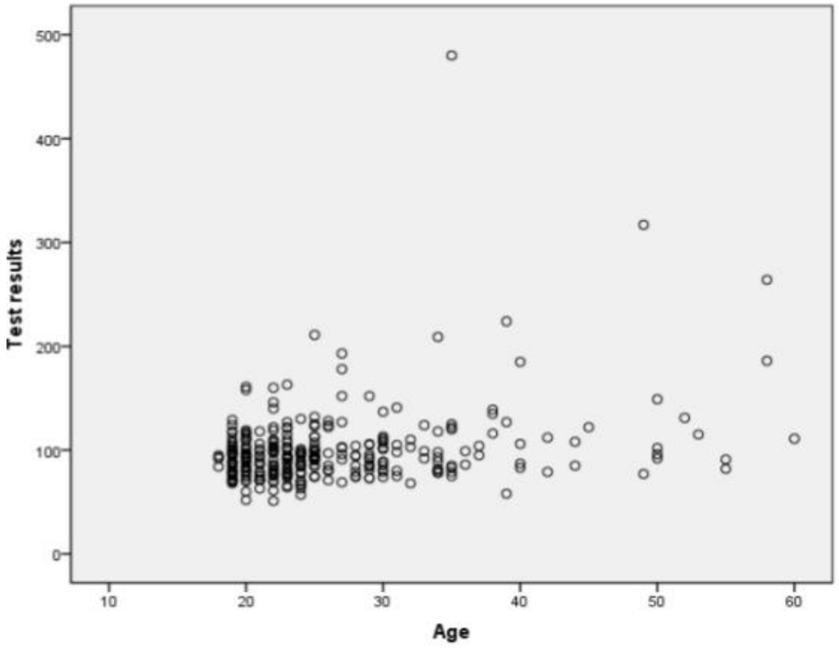
Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear



Source: https://chartio.com/learn/charts/what-is-a-scatter-plot/

Example

Case Study – An analysis that was conducted for diagnosing the presence of diabetes at a workplace



- The population is generally young (75.8% are below thirty).
- This scatter plot illustrates that there is no obvious relationship between age and glucose levels.
- High glucose levels are found in all ages above twenty, and normal glucose levels are found in higher ages.

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation - Barchart

Bar chart

Bar chart / Bar Graph represents categorical data with rectangular bars.

Each bar has a height corresponds to the value it represents.

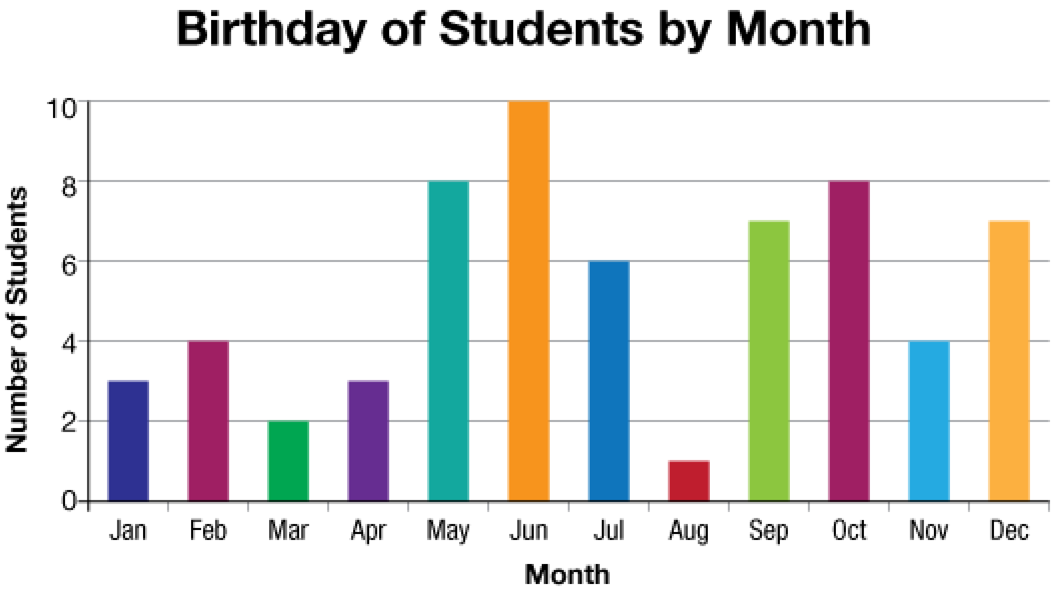
It’s useful when we want to compare a given numeric value on different categories.

Bar chart is to categorical data as histogram is to measurement data.

Bar charts are great when we want to track the development of one or two variables over time.

Let’s understand Bar Chart

Rohan wanted to arrange the birthdays of his classmates according to their months. He then wants to display the data in a way that he can visually understand them. This is where he faces a problem. He doesn't want to use a pictograph since it takes too much time also might be difficult. He want to do it in an easier method. This is where bar graph comes in.



Source: https://www.cuemath.com/data/bar-graphs/

How Bar Chart is different from Histogram?

Histogram

Histogram bars are supposed to touch each other (no spaces between bars) since they represent a continuous variable.

Show distributions of variables.
Display frequencies of ranges (intervals, bins).
Will appear different for different bin sizes.

Bar Charts / Graphs

Bar graphs/charts provide a visual presentation of categorical data.

Bar graphs on the other hand have spaces between them, since bar graphs are used for qualitative (discrete) variables.

How Bar Charts will look like?

Bar charts can be drawn vertically and horizontally, arranging the bars accordingly.

A basic bar chart is drawn vertically and is called a vertical bar graph.

A horizontal bar chart is used when there are a lot of categories involved and they can’t be fitted into the horizontal axis.

Vertical bar charts are also referred to as line graphs.

The x-axis in this represents the data being compared and the y-axis denotes the measured value.

Source: https://www.pluscharts.com/bar-charts/

Features of Bar Chart

A bar chart represents data categories using vertical or rectangular bars that are proportional to numerical values.

It highlights the relationship between data groups and statistical values.

A bar graph details changes in data groups over time.

A bar chart shows the frequency of each data category.

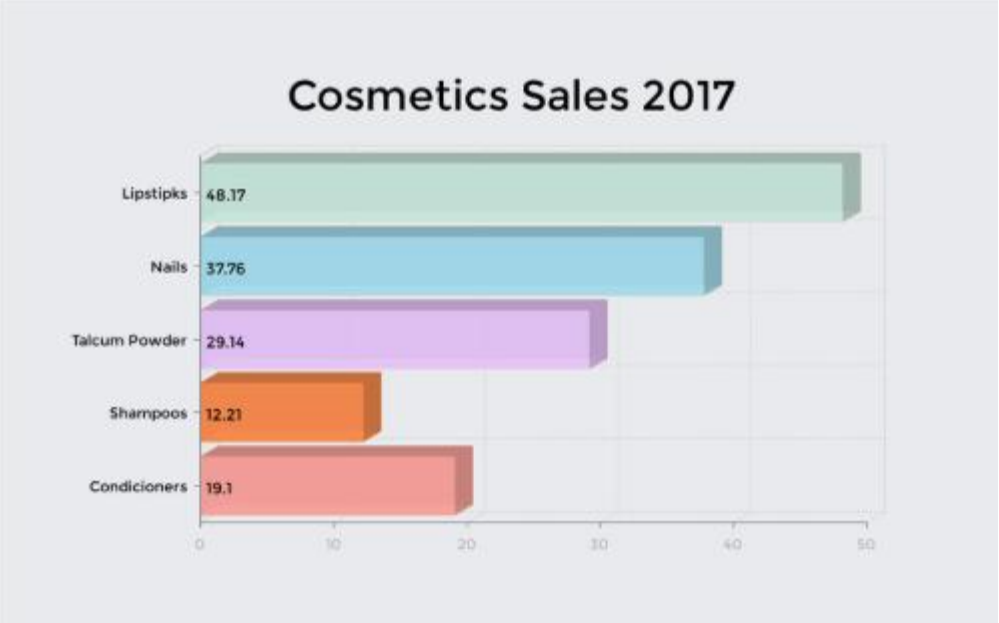
Source: https://formpl.us/resources/graph-chart/bar/

Why Bar Charts are Important?

The importance of bar graphs come out when we want to compare data sets that are independent from one another.

For example, sales report graphs for different quarters or years.

They help in analyzing patterns over long periods of time and lend great value to predictive studies.



Source: https://www.cuemath.com/data/bar-graphs/

What type of Data can be recommended for Bar Charts?

Two types of data are conveyed through bar charts, namely; nominal and ordinal.

Nominal data consist of descriptive data, that aren’t ordered, which provides information regarding an event or a group.

For example, the subject studied at a university.

Ordinal data unlike nominal data is in an orderly fashion and can include different categories too.

For example, a restaurant taking a survey from its customers about their service.

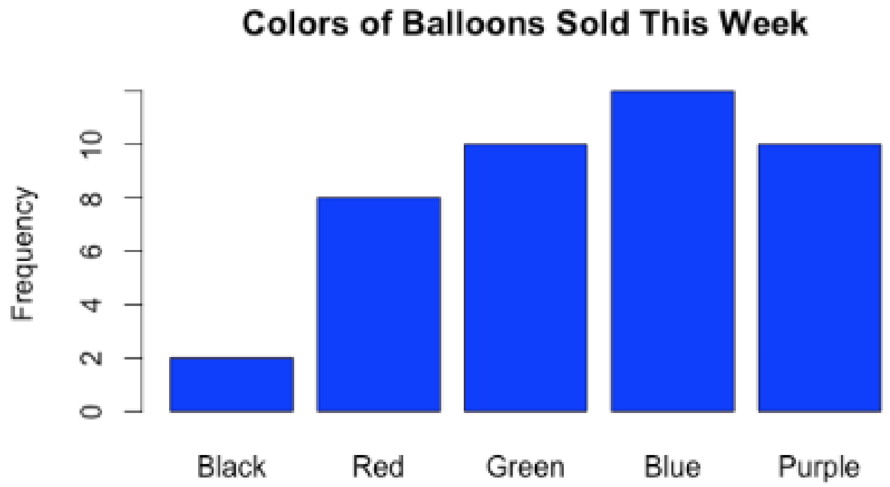
Source: https://www.pluscharts.com/bar-charts/

Categories of Bar Chart – Simple Bar Charts (Vertical)

Typically, a bar graph is drawn vertically and the taller the bar is the larger the category is.

The classes are displayed on the x-axis, and the values(scores) of those classes are displayed on the y-axis.

Useful only when comparing one set of data.

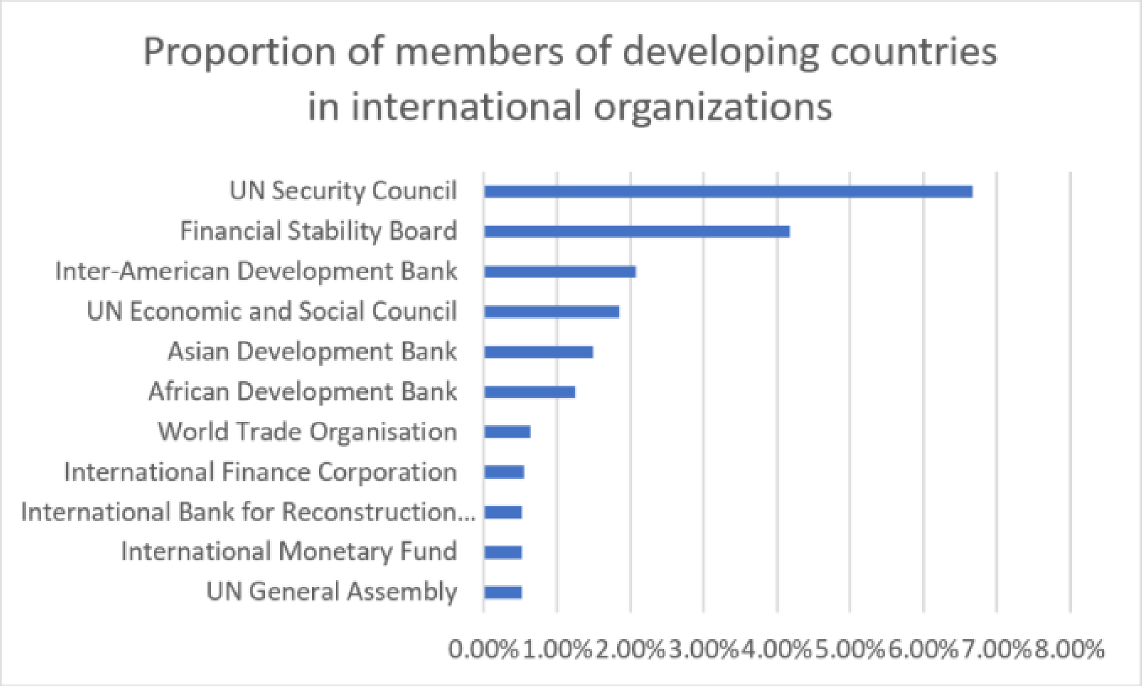


Source: https://www.pluscharts.com/bar-charts/

Image: https://www.stats4stem.org/describing-data-categorical-vs-numerical/

Categories of Bar Chart – Simple Bar Charts (Horizontal)

This is a particularly effective way of presenting data when the different categories have long titles that would be difficult to include below a vertical bar, or when there are a large number of different categories and there is insufficient space to fit all the columns required for a vertical bar chart across the page.



Source: https://www.pluscharts.com/bar-charts/

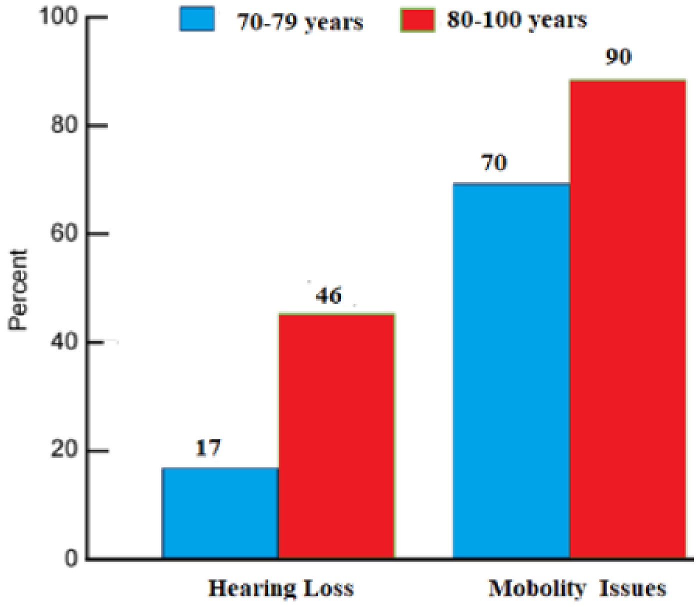
Image: https://stephanieevergreen.com/graph-text-should-be-horizontal/

Categories of Bar Chart – Grouped Bar Charts

Grouped bar charts are known by many names, column bar charts, multi-set bar chart, clustered bar graph, multi-set bar chart, etc.

This chart type is used to compare categories of different groups and can be drawn vertically or horizontally, based on our needs.

Multiple series can be compared in different categories using this chart. The bars representing each category are arranged side by side, such that the differences in the same group is visual and easy to understand.



Source: https://www.pluscharts.com/bar-charts/

Image: https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/ /

Categories of Bar Chart – Stacked Bar Charts

Stacked bar charts are very similar to grouped bar charts. In this type of chart, sub-groups are displayed that fall under different categories.

The sub-groups are placed on top of each other or side by side to make one rectangular bar.

The height or length of the bar shows the category size that it denotes.

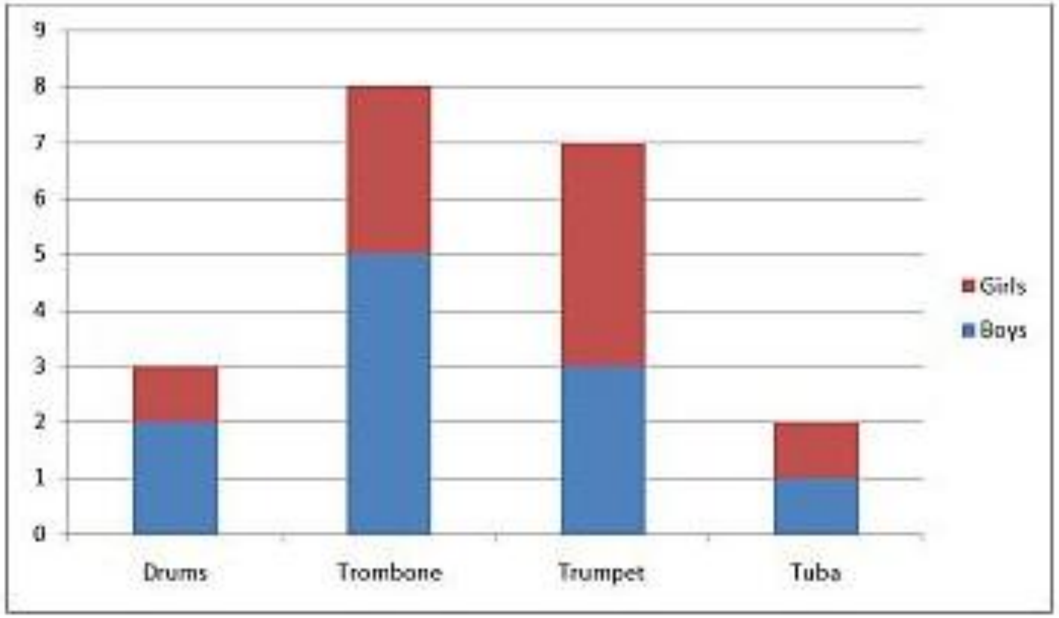
To differentiate between the subgroups, each sub-group is given a different color.

Source: https://www.pluscharts.com/bar-charts/

Categories of Bar Chart – Stacked Bar Charts

For example, take the number of girls and boys in a school learning a musical instrument.

If you wanted to represent the percentage of total students playing the instruments, along with displaying what percent is boys and girls, a stacked bar chart can be used.



Source: https://www.pluscharts.com/bar-charts/

Benefits of Bar Charts

1. Bar charts are made of simple rectangular bars, making it easier to draw them.
2. The scales and figures are easy to read.
3. It shows all the categories present in a distribution.
4. Bar charts summarize large complex data into an easy visual format for understanding.
5. Changes or differences in groups are easy to point out than in tables.
6. Estimations and Calculations can be easily made using this chart.
7. Bar charts are most effective with discrete data, such as rainfall over a month.

Source: https://www.pluscharts.com/bar-charts/

PES UNIVERSITY ONLINE

STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation – Heat Map

Heat Map

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors.

Data points are defined by an x and y axis intersection and a third value that determines the data point's color.

A heatmap is very useful in visualizing the concentration of values between two dimensions of a matrix.

This helps in finding patterns and gives a perspective of depth.

Source: help.highbond.com
www.dundas.com

Heat Map

Heat Maps are extremely versatile and efficient in drawing attention to trends, and it’s for these reasons they’ve become increasingly popular within the analytics community, but that’s just the tip of the iceberg as to why.

While other data visualizations must be interpreted – either by analysts or business users – Heat Maps are innately self-explanatory.

The darker the shade, the greater the quantity (the higher the value, the tighter the dispersion, etc.).

When existing data visualizations are paired with Heat Maps, their ability to rapidly communicate key data insights to the viewer is greatly enhanced.

Source: www.dundas.com

When to use Heat Map?

1) To show a relationship between two measures

2) To illustrate an important detail

3) To use a rating system

4) Use heat maps to compare variables across a large number of categories and to sort complex data by color intensity.

Source: www.klipfolio.com

How does it works?

Data values

Data values appear as boxes on the heat map.

The size and color of each box are determined by the data for that item:

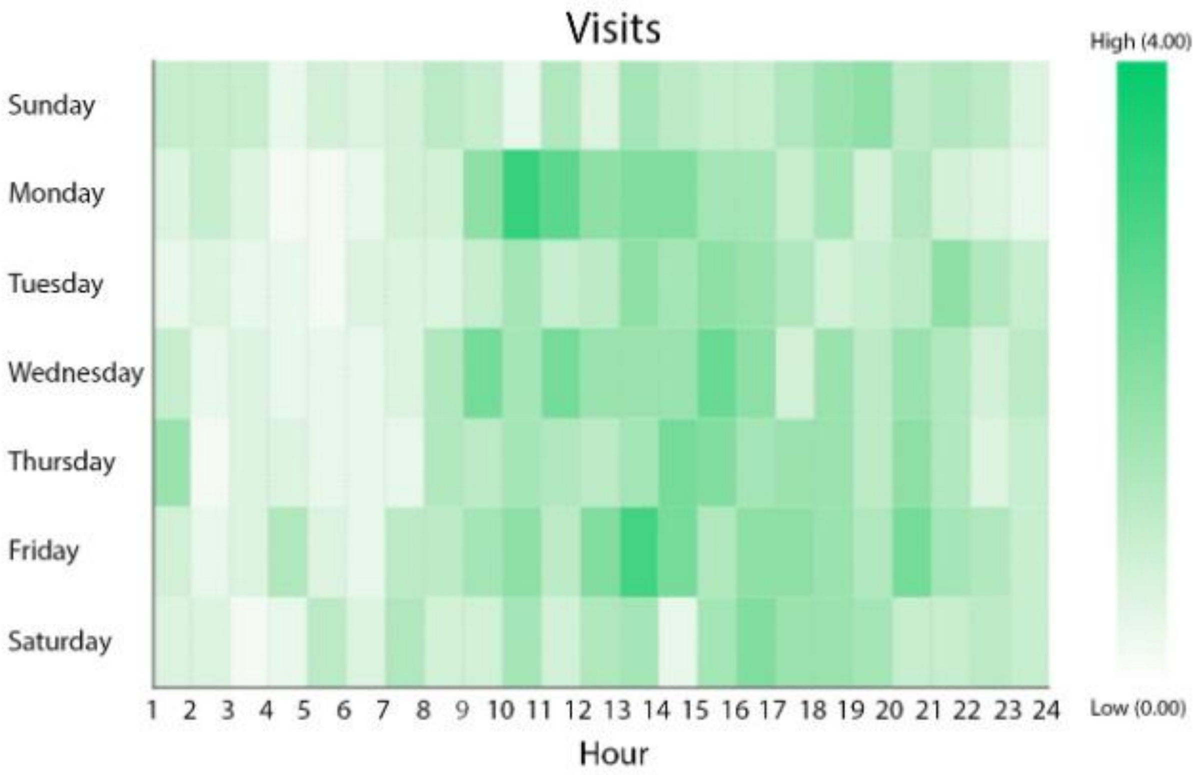
Size - determined by the concentration of x and y axis categories and is not configurable.

Color - determined by the calculated value specified in the **Color by** setting

Source: help.highbond.com

Example – Sales Store

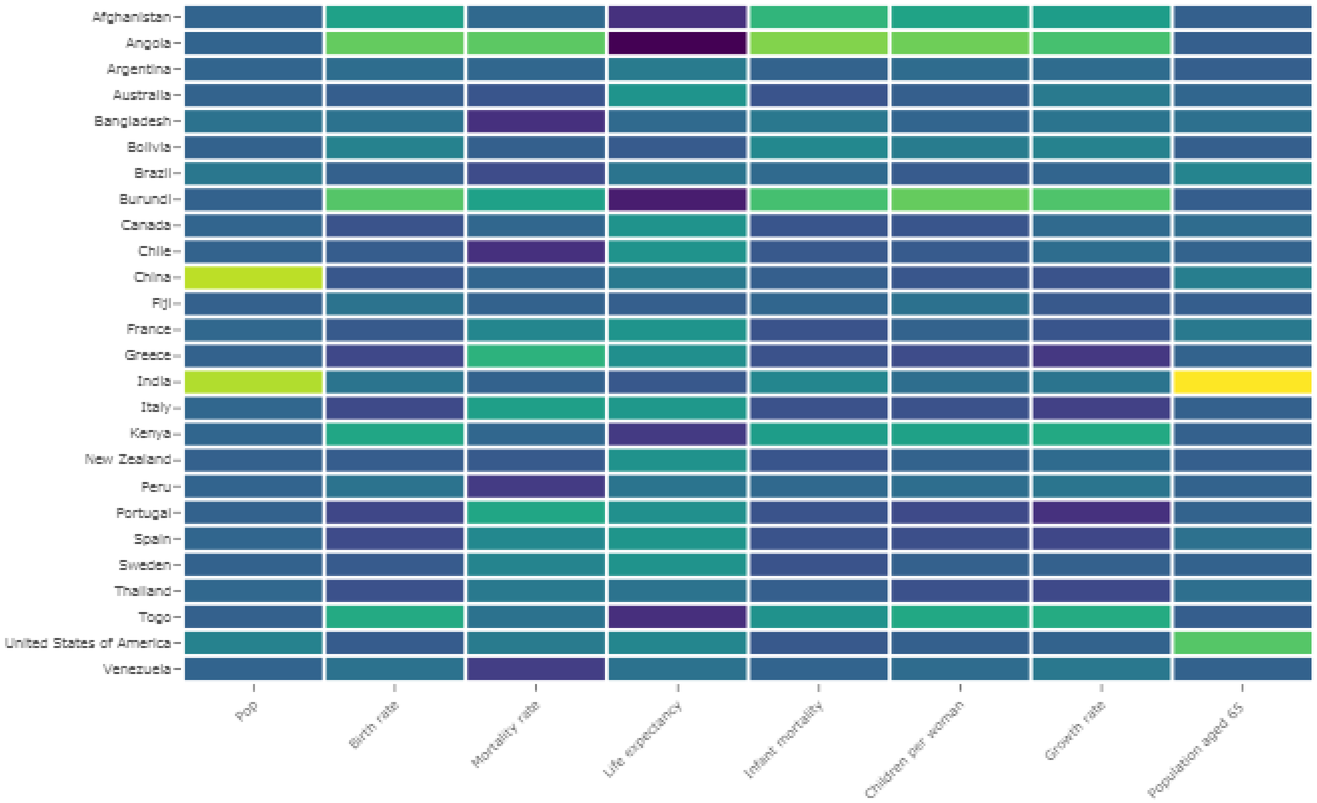
Imagine working as an analyst for a large, multi-national retail corporation that operates a chain of large department stores. To determining whether or not specific dates and times receive more traffic, in order to better allocate in-store resources.



Source: www.dundas.com

Example - Population

Heat map is really useful to display a general view of numerical data, not to extract specific data point. In the graphic above, the huge population size of China and India pops out for example.



Source: https://www.data-to-viz.com/graph/heatmap.html

PES UNIVERSITY ONLINE

THANK YOU

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering