ReDI School of Digital Integration

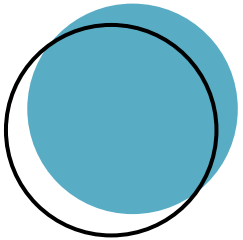
Data Visualization

Lecture 4

We use tech to connect human potential and opportunity with dignity & humility

Learning Objectives

- What is Data VIsualization?
- Plot Types
- Where and when to use?
- Timestamp object
- Choropleth maps



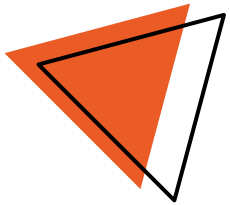
Data Visualization

Purpose

“Data visualization is the graphical representation of data and information.”

“Data visualization is used to identify patterns and trends, outliers, reveal relationships between different variables, and communicate insights and findings to a broader audience.”

It involves the use of charts, graphs, maps, and other visual elements to communicate complex data in a clear and concise manner.

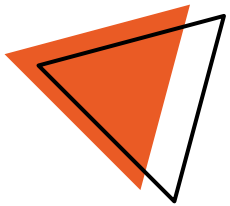


Goal of data visualization:

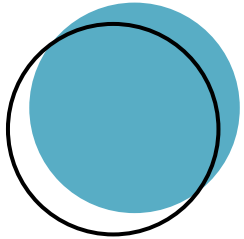
The goal of data visualization is to make data **more accessible, understandable, and useful to users**, leading to better decision-making, improved communication, and more informed insights.

By using data visualization techniques, businesses, organizations, and individuals can make **informed decisions.**

Final words on data Visualization is, it enhances understanding of complex data, improves communication and saves time, facilitates exploration and provides context and perspective, leading to better decision-making and more informed insights.



Types of Data Visualization



- Basic Charts (line, bar, pie, scatter)
- Advanced Charts (heatmaps, treemaps, radar charts)
- Geospatial Visualization (maps, choropleth maps)
- Network Visualization (flow charts, network graphs)

Design Principles

Effective data visualization requires careful consideration of design principles, including color theory, typography, layout, and visual hierarchy.

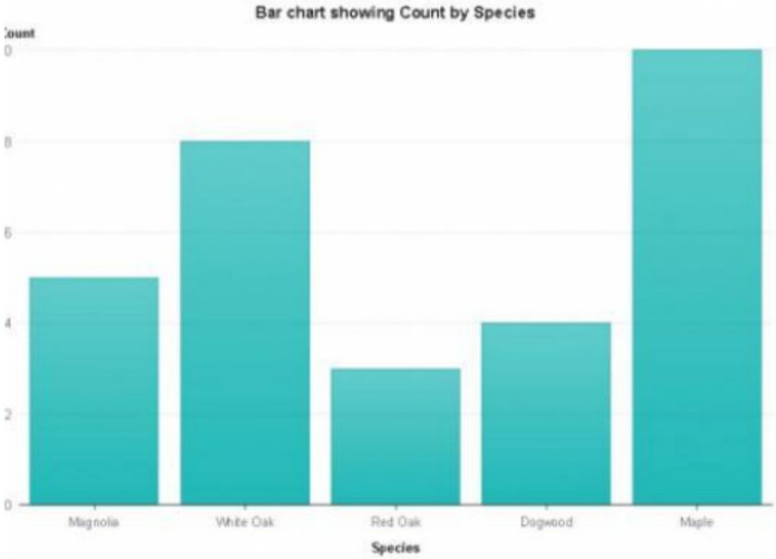
Color Theory: Color can be used to convey meaning and it is important to choose a color palette that is accessible to all users. For example, colors that are easily distinguishable by those with color vision deficiencies should be chosen. Additionally, colors should be used consistently and in a way that supports the overall message of the visualization.

Typography: Font choices can impact readability, and it is important to choose fonts that are legible and appropriate for the context. Fonts should be used consistently and sparingly, and they should be large enough to be easily read.

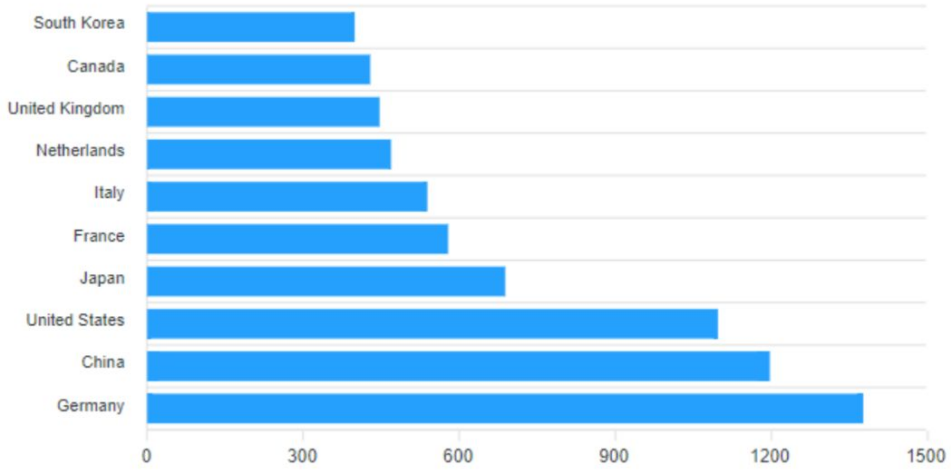
Layout/Visual Hierarchy: The size, color, and placement of visual elements can be used to guide the viewer's attention and emphasize key points. For example, larger elements can be used to indicate importance, and contrasting colors can be used to create visual interest.

Let’s get started!

THE BAR / COLUMN CHART



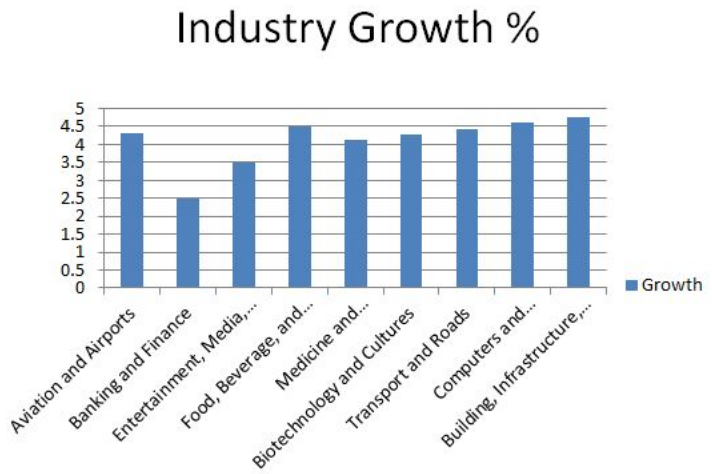
Column Chart

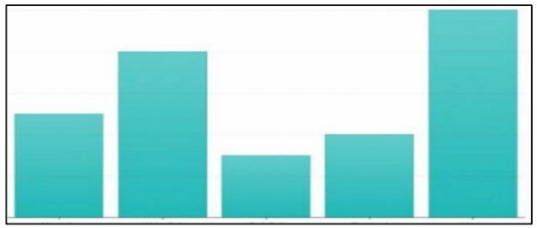


Bar Chart

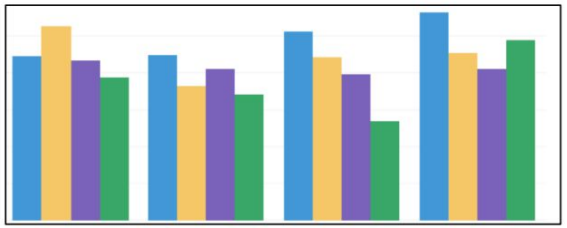
Design aspects:

- To compare Qualitative variable with the Quantitative value assigned with the variable.ex:-Box plot
- Show the whole Y-axis
- Sorting out the bars based on their values.
- Space between the bars.
- Rotating the X-axis to 45 degrees.

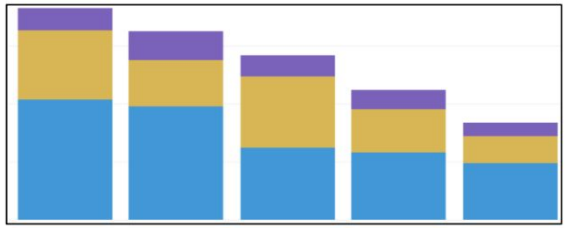




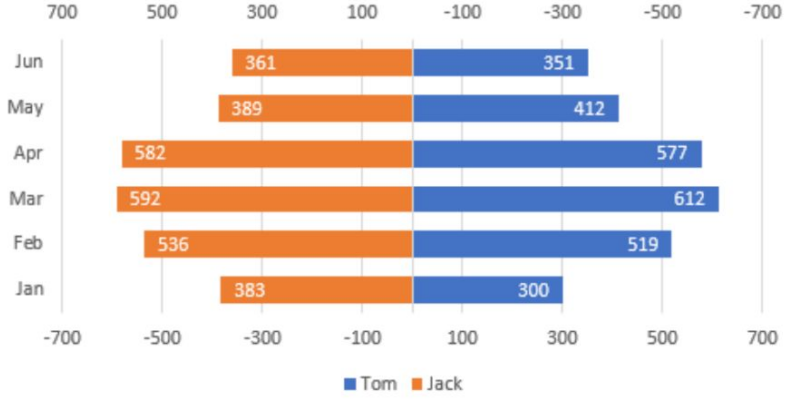
Column Chart

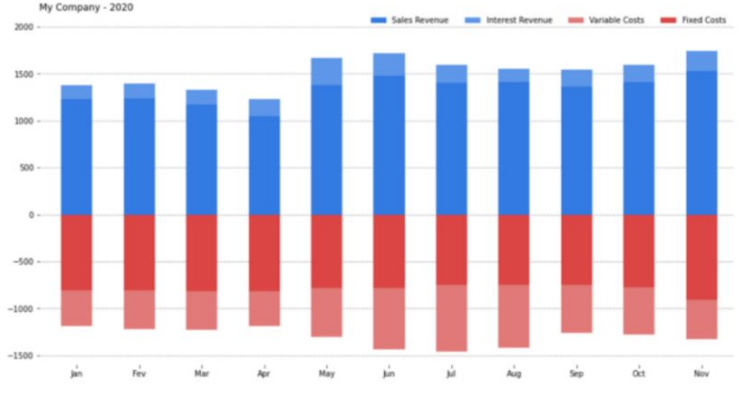


Grouped Column Chart



Stacked Column Chart

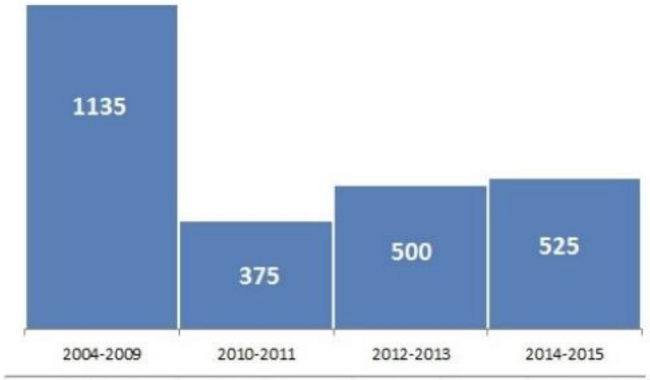


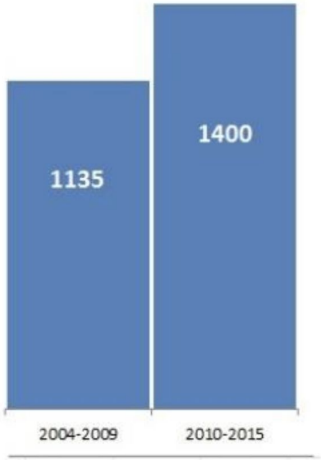


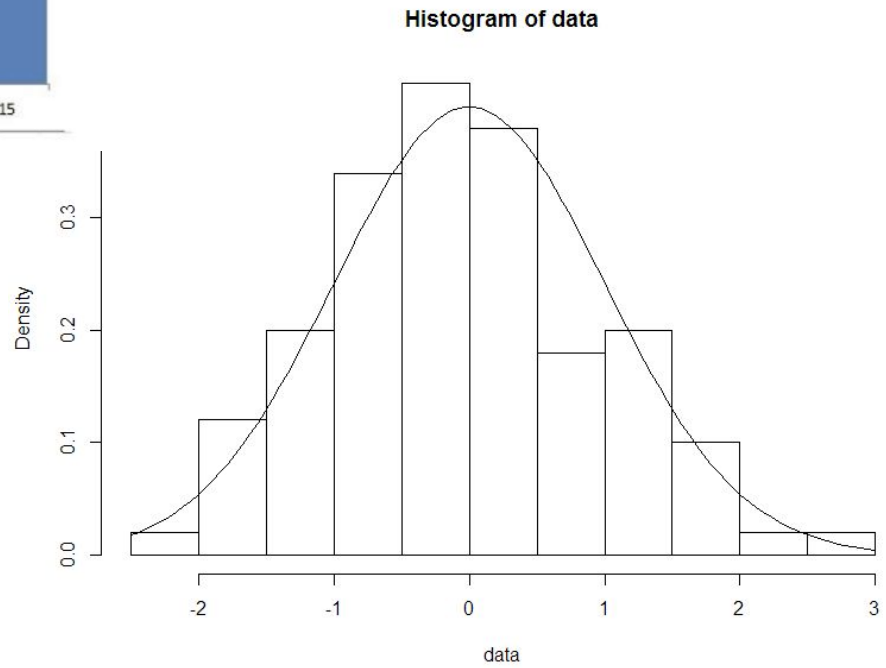
Histograms

Histogram

- A histogram is a graph that shows the frequency of numerical data using rectangles.
- The height of each bar represents the number of values in the data set that fall within a particular bin. When the y-axis is labeled as "count" or "number", the numbers along the y-axis tend to be discrete positive
- The bar graph is the graphical representation of categorical data. A histogram is the graphical representation of quantitative data.

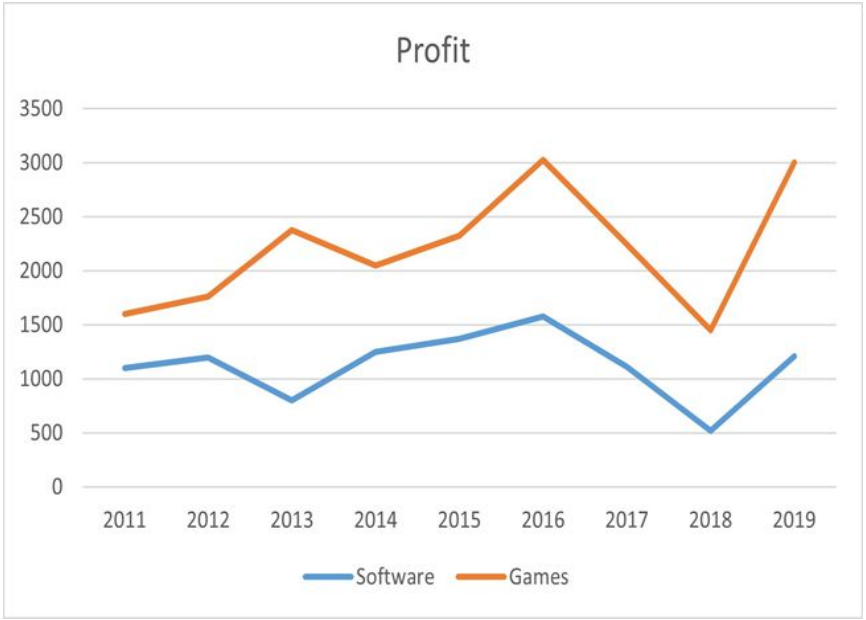

1135
375
500
525
2004-2009
2010-2011
2012-2013
2014-2015


1135
1400
2004-2009
2010-2015


Histogram of data
Density
0.0
0.1
0.2
0.3
-2
-1
0
1
2
3
data

Line Plot

Line Plot



Line Plot:

1.Ideal for time series data or data with a clear order.

2.Shows trends and patterns over time or across a sequence.

3.Connects data points with lines to illustrate relationships between them.

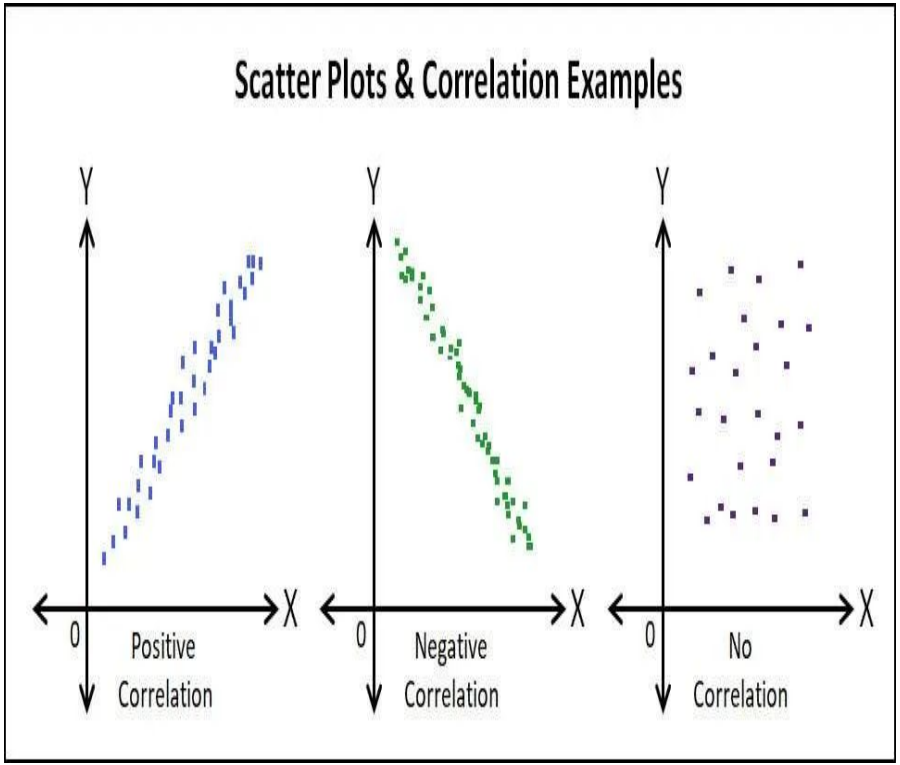
4.Can be used to compare multiple data sets or categories over time.

5.Best for showing continuous data and smooth transitions.

LIMITATION: Can't handle large amount of data!

Scatter Plot

Scatter Plot



Scatter Plot:

1.Ideal for comparing two variables, usually in 2D plots.

2.Shows the distribution and correlation between the variables.

3.Displays individual data points without connecting them with lines.

4.Can be used to identify outliers or clustering of data.

5.Best for showing relationships or lack thereof between two discrete variables.

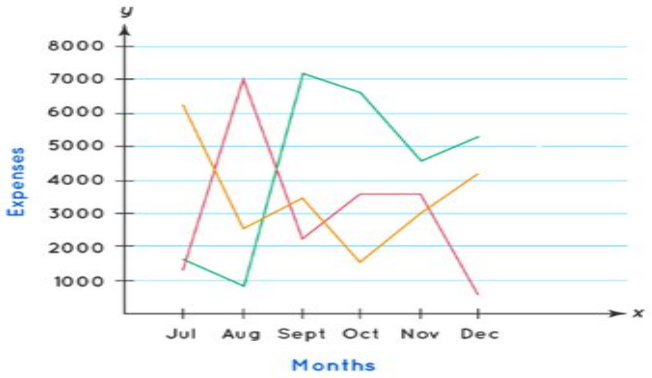
LIMITATION: shows the correlation between the variables but not much emphasis on the exact values.

Line Plot vs Scatter Plot

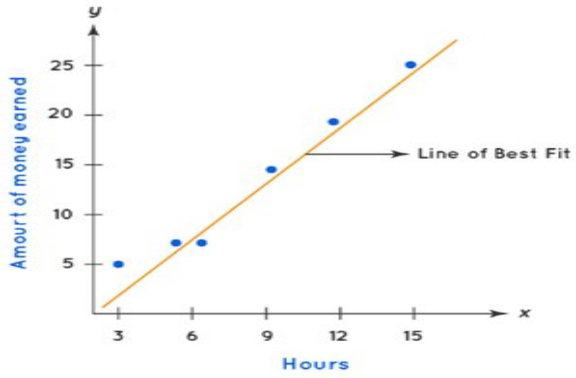
Line Plot vs Scatter plot

We can know the trend of growth by looking at the line in a line graph that connects the data points, whereas in a scatter plot a line of regression or the **'line of best fit'** is drawn which depicts the relationship between two different sets of data along the x-axis and the y-axis.

The **key difference** between a line graph and a scatter plot is that line graph gives how quantitatively the data along the y-axis increases with respect to a given time period, whereas, in a scatter plot, we can see how the data represented in the y-axis changes with increase in the value of data in the x-axis.



Line Graph



Scatter Plot

Pie Chart

Pie Chart

“A circular chart divided into sectors, where each sector represents a proportion of the total. Used to display the relative sizes of data categories or percentages in a whole. “

Pie Chart

When to use Pie Charts:

Comparing relative sizes of a few categories (usually no more than 5-7).

Representing percentages or proportions in a whole.

Showcasing a simple breakdown of categorical data.

Visualizing data where categories have distinct differences.

Disadvantages of Pie Charts:

Difficulty in comparing the size of different sectors, especially if they are similar in size.

Ineffective for displaying multiple data series or a large number of categories.

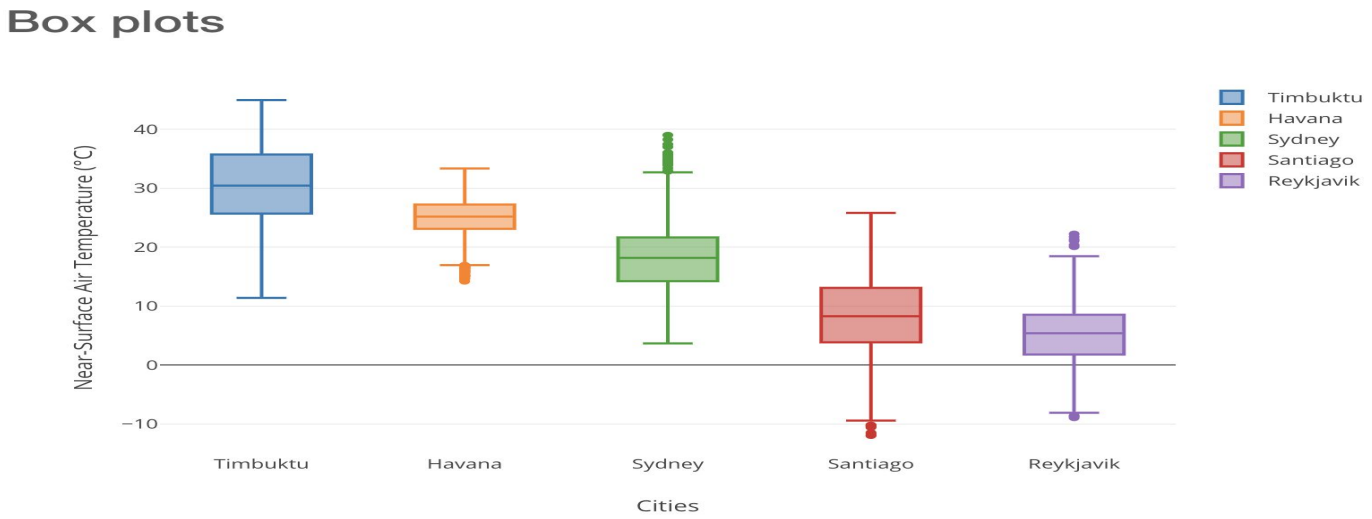
Can be misleading if the data is not properly represented or if the segments are not labeled. Difficult to accurately perceive small differences between sectors.

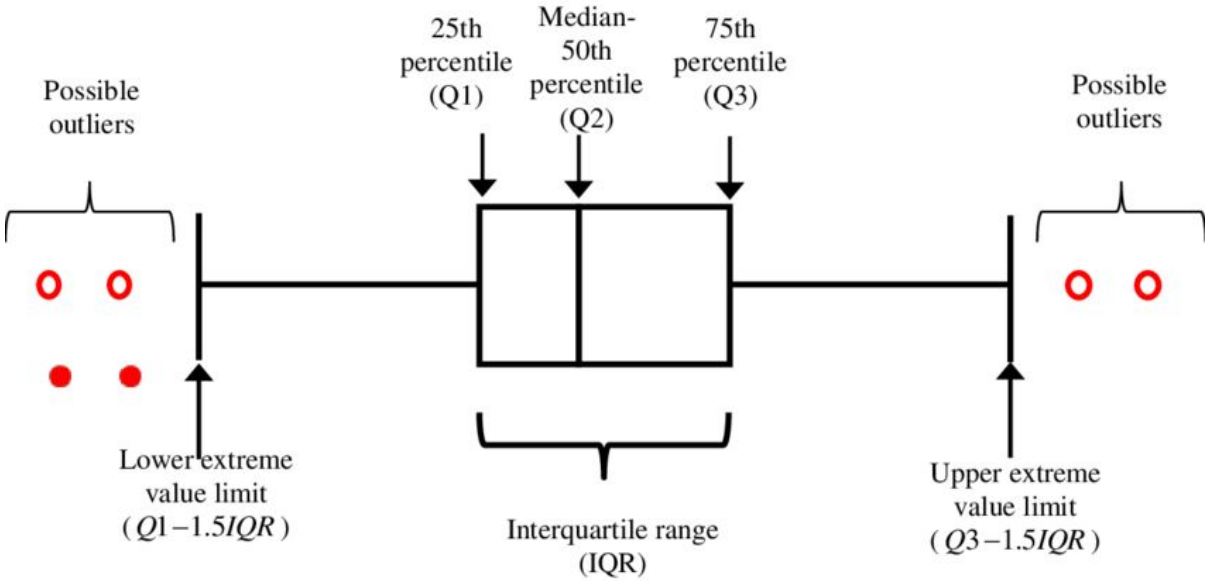
Limited ability to show trends or changes over time.

Box Plot

Box Plot

A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset's five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. It displays the central tendency, spread, and skewness of the data, and can also reveal outliers




Possible outliers
25th percentile (Q1)
Median-50th percentile (Q2)
75th percentile (Q3)
Possible outliers
Lower extreme value limit (Q1 − 1.5IQR)
Interquartile range (IQR)
Upper extreme value limit (Q3 − 1.5IQR)

Box Plot

Advantages of Box Plots:

Compact representation of a dataset's distribution and spread.

Easy comparison of multiple datasets side by side.

Identification of outliers and potential data entry errors.

Can visualize skewness in the data.

Not influenced by extreme values, as the box only covers the interquartile range (IQR).

Disadvantages of Box Plots:

Limited detail about the dataset's shape, as it only shows summary statistics.

No information about the underlying frequency distribution.

Can be misleading for small sample sizes or non-symmetric data.

Difficulty in interpreting for non-experts without context.

Time Series Visualization

Time Data: *"Time series data refers to a sequence of data points collected or recorded at regular time intervals. It is a type of data that is ordered chronologically and can be used to analyze trends, patterns, or seasonality in the data over time."*

```
ts_pd=pd.to_datetime(ts, format="%Y-%m-%d %H:%M:%S")
print(ts_pd, "\n", type(ts_pd))
```

```
df["timestamp"]=pd.to_datetime(df["timestamp"], format="%Y-%m-%d %H:%M:%S")
type(df.loc[0,"timestamp"])
```

Format

%m	Month as a zero-padded decimal number.	01, 02, ..., 12	(9)
%y	Year without century as a zero-padded decimal number.	00, 01, ..., 99	(9)
%Y	Year with century as a decimal number.	0001, 0002, ..., 2013, 2014, ..., 9998, 9999	(2)
%H	Hour (24-hour clock) as a zero-padded decimal number.	00, 01, ..., 23	(9)
%I	Hour (12-hour clock) as a zero-padded decimal number.	01, 02, ..., 12	(9)
%p	Locale's equivalent of either AM or PM.	AM, PM (en_US); am, pm (de_DE)	(1), (3)
%M	Minute as a zero-padded decimal number.	00, 01, ..., 59	(9)

Timestamp Object in Pandas

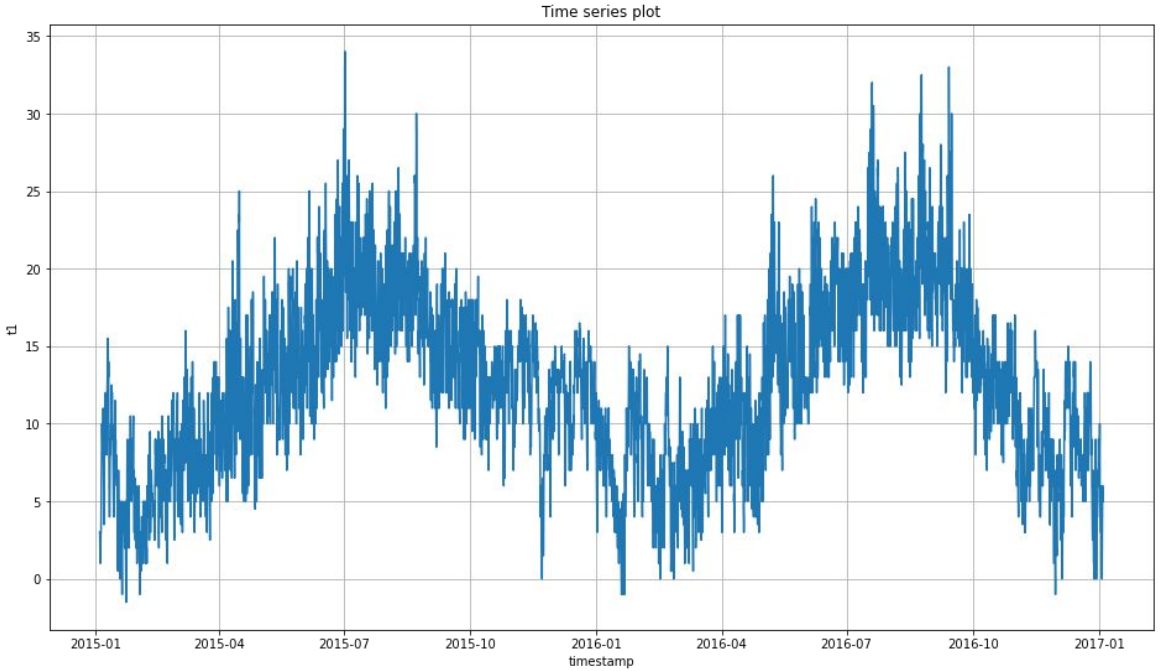
In Pandas, a Timestamp object is used to represent a single date and time. Timestamp objects can be used to index Pandas data structures, such as Series and DataFrame, and can be used to perform various time-related operations.

- Timestamp objects can be used in a variety of ways, such as to index a DataFrame by date or to perform **time-based filtering** of data. For example, we can use a Timestamp object to select all data from a DataFrame that falls within a certain time range.
- **Frequency Conversion:** Timestamp objects can be used to perform frequency conversion, allowing for easy resampling of data to different time frequencies. This can be useful for tasks such as aggregating data to a different time scale or calculating rolling averages.
- **Time Series Visualization**: Pandas includes built-in support for time-series visualization, allowing for easy creation of time-series plots and other visualizations.

```
ts = df["timestamp"][0]
print(ts.date())
print(ts.day)
print(ts.year)
print(ts.weekday())
print(ts.day_name())
print(ts.hour)
```

Time Series Plot

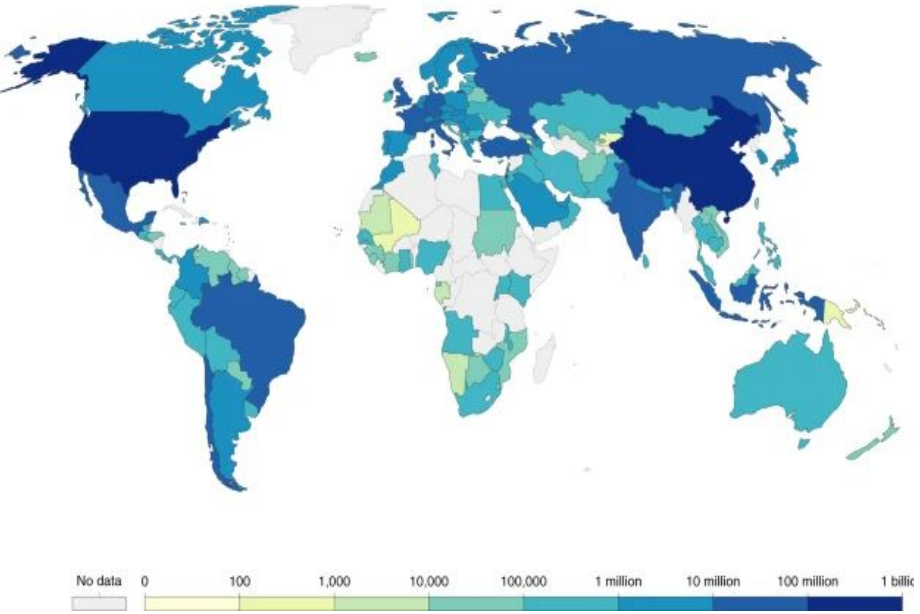
“ A time series plot is a type of data visualization that is used to display changes in data over time. Time series plots are commonly used in fields such as finance, economics, and engineering, as they allow users to identify patterns and trends in data over time.“

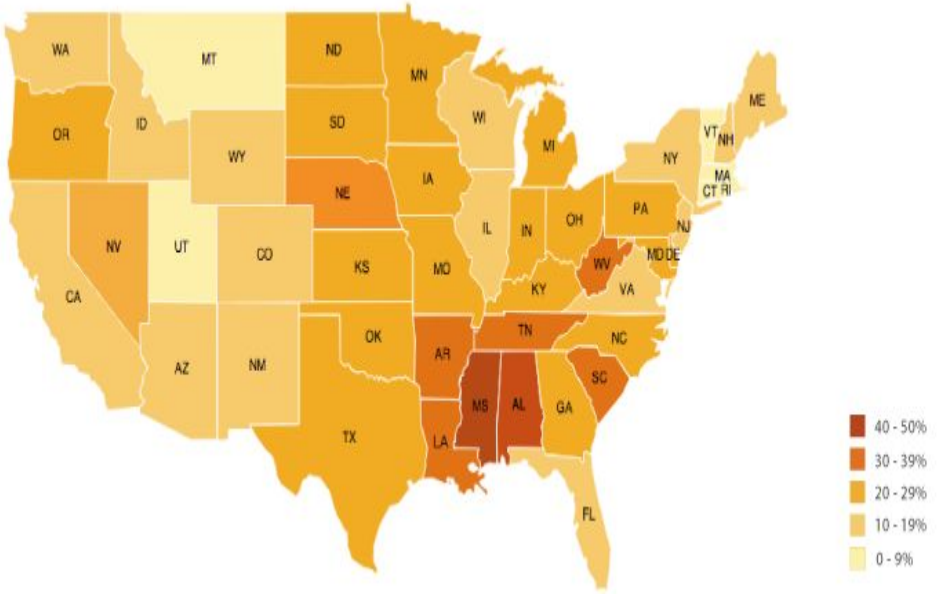


Map Visualization - Choropleth maps

Choropleth map

“Choropleth maps are commonly used to visualize data that varies across geographic regions, such as population density, per capita income, or election results. In a choropleth map, each region is shaded with a color that corresponds to the value of the data being visualized.“





WE DID IT!