Unit 1

Introduction to Visualization

Data analysis is the process of cleaning, processing and changing raw data and extracting relevant information from it. Its main purpose is to look for trends, grouping and other relationship between different types of data. **Data visualization** is the process of translating information into visual context such as graph, charts, map etc. to make data easier for human to understand and gain insight from it. The aim of data visualization is to make it easier to identify patterns, trends and outlier in data sets. Visualization represents analyzed data in form of pictorial form that helps to communicate results in simpler form. Data visualization eccentrics includes formats like bar charts, line graph, scatter plot, heat maps etc. Data visualization is a way to represent complex information in easier and more understandable format. Effective data visualization are designed to be clear, accurate and visually appealing such that human brain can perceive and understand meaning in easier way.

Good visualization helps to :

- Provide rapid access to data
- Represent the data in proper format and tell the scenario
- Express complex ideas in simple form
- Shows relationships between abstract concepts

Visualization stages:

it generally includes four basic stages:

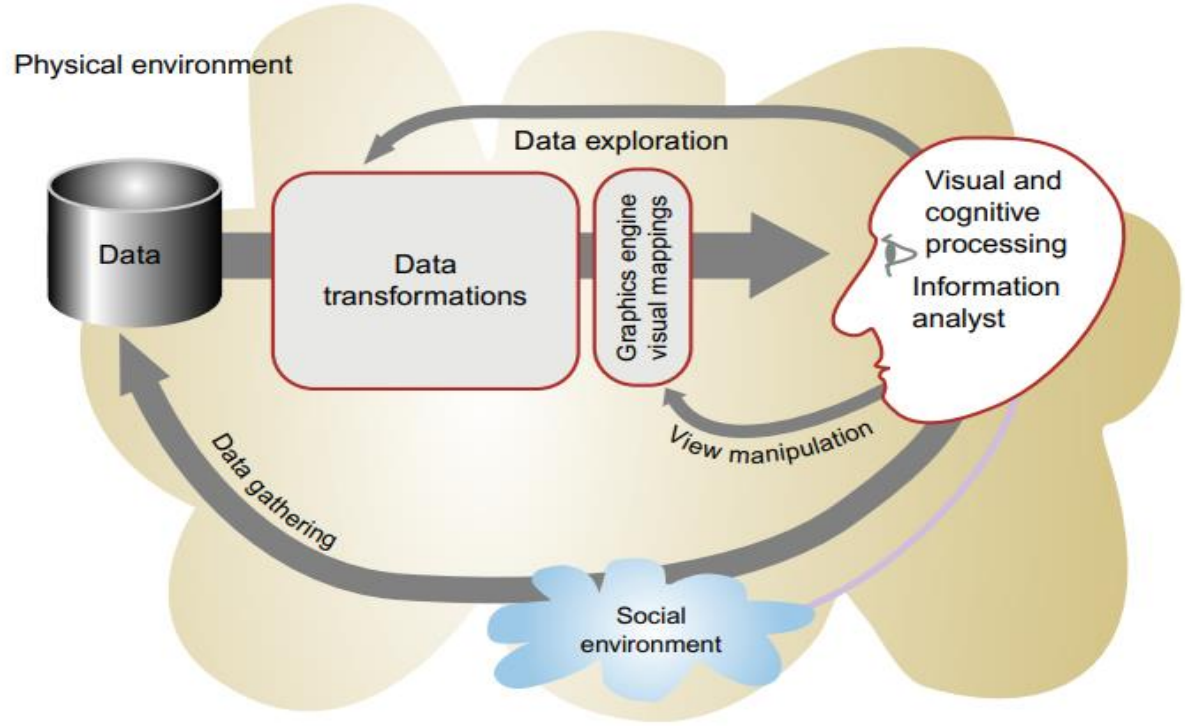
Step 1: collection and storage of data

Step 2: preprocessing stage which transform the data into something that is easier to manipulate.

Step 3: mapping from the selected data to a visual representation which is done through the algorithms that produce an image on the screen.

Step 4: the human perceptual and cognitive system.

Visual perception refers to the way in which human perceive and interpret visual information represented by visual eccentrics like charts, graph etc. it refers to how human's visual system process and interpret the surrounding such as color, shapes, sizes to extract meaningful information and pattern from data. It is the ability to interpret the surrounding environment by processing information that is contained in visible light. It explains how people express surroundings through light that enters our eyes. For effective data visualization, visual perception should take into consideration such that accurate and understandable visualization can be created.



In above figure, longest loop involves data gathering. A data seeker may choose to gather more data to follow up on an interesting lead. Another loop controls the computational preprocessing that takes place prior to visualization which helps to give up the meaning form subjected data. Both the physical environment and social environment are involved in the data gathering lop. The physical environment is a source of data while the social environment determines the way the data is collected and how it is interpreted. The computer is treated as a universal tool for producing interactive graphics i.e. once we figure out the best way to visualize data then algorithm is constructed to create appropriate images.

A visual representation of data refers to use of graphical or visualization tool to convey patterns, information, relationship in dataset in easier and simpler form. The main goal of visual representation is to enforce effective communication and analysis of information. The choice of visual representation depends on the type of data, meaning we want to convey and the audience familiarity with the data domain. Some common types of visual representation are:

1. **Charts and graphs:** includes bar chart, line charts, scatter plot, pie chart that are used to display quantitative data to reveal trends and comparison.
2. **Maps**: use to represent geographical data using heatmaps, maps and other spatial visualization. It helps to reveal geographic pattern and their relationship
3. **Tree map:** use nested rectangle or circular segment to represent hierarchical data and its distribution within categories.
4. **Heatmap**: use to represent two-dimensional data by using color to show density of data points and highlight patterns and concentration.
5. **Word clouds:** use to display words from a text dataset with the varying size of text to indicate its frequency. It helps to visualize word frequencies and marks the important topic in text analysis.

6. **Time series visualization:** shows how data change over time and includes charts to visualize chart.

Data abstraction is the process of summarizing and simplifying complex data to highlight required patterns, trend and insight. It is a crucial process to make data more manageable and understandable for data analysis and visualization pattern. It is used to reduce level of details from data set. Effective data abstraction helps to enhance the efficiency of data by focusing on most relevant aspect of data. Data abstraction is required in data analysis and visualization for:

i. **Reducing complexity:**
Complex data sets contain null values, noise and irrelevant details. So data abstraction helps to filter out noise, null values and focus on the important aspect of data. Simplification in data makes easier to identify pattern and relationship between variables.

ii. **Aggregation:**
Involves combining individual data points into group to calculate summary statistics of each group. For example: daily sales figure, population trends, disease trend etc. aggregation helps in understanding overall trends and patterns in each data points.

iii. **Summarization:**
Involves creating lot of information to capture essence of the data sets. It includes calculation measures like mean, median, mode, range and percentile that can helps to provide quick overview of the data distribution.

iv. **Data transformation:**
Includes operation like scaling, log transformation and normalization which can improve the distribution and relationships within the data to make it more suitable for data analysis and visualization.

v. **Sampling:**
Sampling is used to select a representative subset of data for analysis. This helps to reduce computational and time requirements.

Visual encoding refers to process of mapping data attributes to visual eccentrics in order to create meaningful information and graphical representation. It involves choosing appropriate visual actor like position, color, size shape and texture to represent different aspect of data. It helps to effectively communicate data insights and patterns to the user. Effective visual encoding ensures that user can quickly and accurately understand the patterns, trend and relationship present in the data. Some common visual properties used in visual encodings are:

i. **Position;** the position of axis in different visual eccentrics like scatter plot, bar graph etc. can represent two different variables.

ii. **Size:** size of graphical elements like circle, bar can be used to represent quantitative values. Larger size typically indicates larger values.

iii. **Color:** color can represent wide range of information and distinguish the information like red color indicates danger, green color indicates safe. Color should be used carefully to ensure accessibility and avoid misinterpretation.
iv. **Shape:** different shape like circle, triangle, square etc. and be used to distinguish categories or represent data points.
v. **Texture:** used to differentiate elements in a visualization when color or shape is limited
vi. **Opacity:** the level of opacity can be used to show density or emphasize specific data points.
vii. **Connection:** lines or links connecting elements that can represent relationships or connection between data points. Mostly used in network visualization.

Color in visualization:

Color is a tool to convey information, highlight patterns and engage viewers. In helps to give meaning on datasets. It should be used carefully as it can convey different meaning of data. Color plays the important role on ensuring correct visualization. Color is valuable tools which can enhance the clarity and impact of your data. Color are used to visualize following data:

i. **Categorical data**. Color is used to differentiate groups in dataset. For example, color can be used to represent different product categories, population categories in bar graph, comparison of two or more data in scatter plot, distinguish countries on a map.
ii. **Sequential data:** color can be used to represent ordered data such as values that increase or decrease progressively. For example, different color used in heatmaps to show similarities and dissimilarities between variables, color in line chart to distinguish two or more variable.
iii. **Diverging data**: color can be used to point the data that are above or below a central value. It helps to show distinct midpoint such as comparing positive and negative changes.
iv. **Data density:** color can be used to indicate density of data points in specific area. Lighter or dark shades may represent area of higher or lower concentration.
v. **Time series:** different color can be used in line chart or area chart to represent converging and diverting time fluctuation in dataset that can help user to track trend and fluctuation
vi. **Emphasis and highlighting:** color can be used to draw attention to specific data points or area of interest in data visualization. Highlight data can helps to understand main insights.
vii. **Error:** color can be used to represent error interval, range on data points. Lighter color can indicate higher error and uncertainty whereas dark color can represent more emphasis on data points.

Following points should be taken into consideration while choosing appropriate color in visualization:

- Choose a color palette that is appropriate for the data type and message of the visualization
- Avoid using too many colors as this can lead to confusion and convey different information
- Maintain color consistency across related visualization to make easier for comparison

Perceptual issues:

Perceptual issues are those that can lead to misinterpretation of data due to perception of human brain i.e., it refers to challenges and considerations related to how human perceive and interpret visual representation of data. These issues should be addressed as effective data visualization depends on how well it conveys information to the viewer. Addressing perceptual issues in data visualization requires a deep understanding of both the data being presented and the target audience. It is important to continually test and refine visualization to ensure that they effectively convey the information and minimize misinterpretation. Some of the common perceptual issues are:

- Color choice: the human eye is more sensitive to some color that others so using such similar color and dissimilar color together can make difficult for user to understand the meaning of data. Using inappropriate color can lead to confusion of data for eg. Using red color and green color together for danger symbol, choosing a color scale that is not perceptually uniform can also distort the representation of quantitative data.
- Size and position: using size too similar for different situation can make it difficult to see the difference in data values. Distortion in scale such as unequal axis, intervals exaggerated size in chart and graph can misrepresent the data and lead to inaccurate interpretation. Placing the data points in unexpected or cluttered location can make the information difficult to understand.
- Misleading visualization types: choosing wrong types of visual eccentrics like chart and graphs for data can result in misinterpretation. For example: using pie chart to show time series data, using heat map to represent one dimensional data etc.
- Labeling and annotation: improper labeling and annotation can misinterpret the meaning of data and leave viewers unsure about the context.
- Data density: too much information can overwhelm the user while too little can lead to miss the main points on decision making. So, how much data is to represent should be clearly analyzed.
- Complexity: human eye can only process limited amount of visual information so using visualization that are too complex can make it difficult to understand the data.

To overcome perceptual issue, following points should be taken into consideration:

- Using limited number of colors
- Using sizes that are clearly differentiated
- Using familiar shapes that are not distorted
- Keeping visualization simple and easy to understand
- Keeping limited amount of data for visualization.

Information Overload:

Information overload is the situation where the amount of data presented is too much for the user to understand or process which can happen if there is too much data in a single visualization or when multiple visualization is presented together. Information overload can lead to confusion and decrease ability to understand the data. Information overload can lead to following problems:

- The user may become confused and unable to understand data
- The user may not be able to find information they are searching for.
- The user may become frustrated and give up trying to understand data

Following are the elements that can lead information overload:

- Excess data points: includes excess number of points such as categories, series.
- Complex visual elements: includes complicated charts or graph, multiple lines etc
- Redundant information
- Too many variables
- Lack of hierarchy: if all element are treated equally then it can be challenging for viewer to prioritize and understand the most crucial information.
- Real time data: constant updates and information change can overwhelm the user
- Insufficient context: insufficient explanation can leave viewers in confusion.
- Poor organization of data
- Overuse of color: excessive use of color, styles or formatting options can create visual clutter and make it difficult to focus on the data.

Following factor should be consider to address information overload

- Simplify data: reduce the amount of data and focus on most critical variable and insight
- Prioritize information: highlight the most important data points or trends to guide the viewer attention. Use visual cues, annotation to drew attention
- Use limited interaction: provide clear instruction and options for customization
- Provide context: include clear title, labels, annotation, to provide context and guide. Use caption or summary that can provide more insight to viewer.
- Use hierarchy and consistent design