UNIT-5 Big Data Visualization

@prathilkpatil
@[illegible] Patil, pimpri

9] Explain data visualization with the help of example? What are the advantages of data visualization.

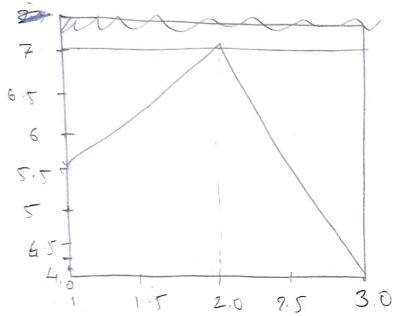
→ • Data visualization is the presentation of quantitative data into graphical form. It helps to turn large and small datasets into visuals that are easier for the human brain to understand & process

- Good visualization are created when data science, communication & design collid.
- Data visualization is the process of translating large data or metrics into charts, graph & other visuals.
- The visual representation of data helps in identify & share real-time trends, outliers & new insights about the information represented in the data.
- In order to make a good visualization, you need to start with clean data & well structured. Once the data is ready to visualize, we need to pick right chart.
- A Graph is simply a visual representation of numeric data. MatplotLib supports a large number of graph & chart types.
- Matplotlib is popular python packages used to build plots.
- To build plots & graphs the matplotlib's pyplot library is ~~instalte~~ imported as plt.

P.T.O.

```
import matplotlib.pyplot as plt
plt.plot([1,2,3],[5,6,7])
plt.show()
```

The plt.plot() will make the plot in background but we want to show it on the screen so we are using plt.show()



Advantages of data Visualization

1) Simplifies complex data
2) Faster Decision making.
3) Identifies patterns & Trends
4] Support real-Time monitoring
5] Easy to understand

9] Explain Challenges in big data Visualization.

→ Big data analytics plays a key role through reducing the data size & complexity. Hence Visualization helps ~~is~~ big data to get complete view of data & discover data values.

- Scalability & dynamics are two major ~~problems~~ Challenges in visual analytics
- Volume: The method are developed to work with large number of dataset & derive meaningful data.
- Variety → The method are developed to combine as many data sources as needed
- Velocity → Helps in real time processing
- Value → Method helps in extracting useful insights or buisness value from big data.
- Visualization → It is difficult to visualize big data because it comes in many different types & formats.

Problems with big data Visualization

1] Lack of skilled Users
2] High performance Required.
3] Information Loss.
4] Visual noise. → most of the data is too related that user cannot divide them as seperate object on the Screen.

5] Data Quality Issues.

- If the data is incorrect or missing, the visual output become unreliable.

6] Wrong Visualization choise.

- Using the wrong chart type can give misleading results or hide the real meaning

Solutions

1) Need for Speed → One solution is to deal with hardware, by increasing memory & massive parallel processing can be used.

2) Understanding the data → Select proper domain and expertise is solution.

3) Displaying meaningful results → Solution is to cluster data into smaller groups that are visible effectively

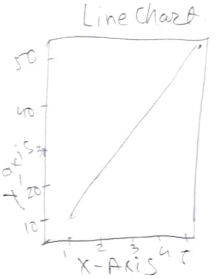
4) Dealing with outliers → Solution is to remove outliers from data or create seperate chart for outliers

8] Write two visualization function from matplotlib

→ import matplotlib.pyplot as plt.

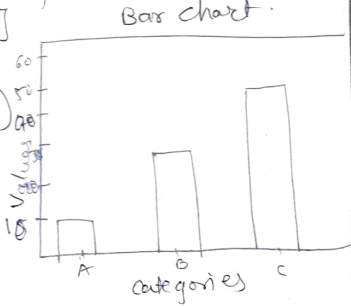
①

```
x = [1,2,3,4,5]
y = [10,20,30,40,50]

plt.plot(x,y)
plt.title("line chart")
plt.xlable("x-Axis")
plt.ylable("y-axis")
plt.show()
```



② import matplotlib.pyplot as plt.

```
categories = ["A", "B", "C"]
values = [10, 30, 50]

plt.bar(categories, values)
plt.title("Bar Chart")
plt.xlabel("Categories")
plt.ylabel("Values")
plt.show()
```



Q] How data visualization helps big data analytics?

→ • Big data visualization is a visual representation of Big data. It could be simple as line chart, bar chart, histogram or pie chart or bit complex like heatmap, scatterplot or tree map.

• Visualization of Big data can also be done in 3-Dimensional graphs. based on the use case.

• Generally when Big data analytics & algorithm are applied to data sets, the results are meant for decision makers. The best part of data visualization tools is that they visualize data without loss of accuracy.

★ Big data visualization Helps in following ways

1) Simplifies Complex data

2) Reveal hidden patterns & Trends.

3) Improves decision making

4) Helps in real time monitoring.

5) Enhances communication → Easy to share with non-technical stakeholders

6) Supports predictive analytics.

Q] Explain any four types of data visualization with example

→ 1) Multidimensional : 2D Area.

These are the most common & basic types of visualization used to represent quantitative data into two axes (X&Y)

1) Cartogram : mainly consists of two main types distance based & Area based

It is special type of map where size of the places is changed to show some kind of data.

eg → showing population density.

2) Choropleth : Area are shaded according to data variables. It is used to represent the statistical measurement such as population density or website visitor ~~per~~ count per city.
→ Average income by region.

3) Dot distortion map : It uses a dot symbol to represent feature on the map.

2] Temporal.

Temporal visualization deals with data over time, showing trends, patterns etc.

1) pie chart → The circle is divided into sectors. to represent numeric proportion.
→ Subject marks.

2) Histogram → In histogram the data are grouped in ranges (eg, 10-19, 20-29) & then plotted as connected bars. Shows frequency of data in continous intervals.

eg → Age distribution, rainfall over months.

3] Scatter plot → Scatterplot is used to find trend, clustering & outliers from a given dataset. Useful when looking for outliers or understanding distribution of data.

3] Hierarchical

- Used to display data structured in level or Hierarchies.

1) dendograms → It is nothing but a tree diagram used to represent clusters
eg → Hierarchical clustering in ML.

2) Ring Chart → It is multi-level pie chart which is represented by the nested circle.

3) Tree diagram: It represents the data or the hierarchy in graph form, which can be visualized from left to right or top to bottom.

4] Network

Shows connections & relationship between entities.

1) Alluvial diagram.

Shows flows & relationships between groups or states over time.

eg → website traffic from source to destination

2) Node-link diagram → In this diagram nodes are represented using dots and link between them is shown as line segments to display the data connections.

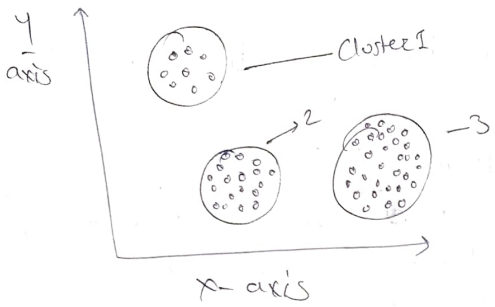
eg.

Q] Explain scatter plot, histogram & heat map with example.

→ Scatter plot:

- Scatter plot is used to find trends, clustering & outliers from a given dataset.
- This is useful when looking for outliers or for understanding the distribution of your data.
- The scatterplot uses x & y axis to show the data and it shows the relationship between two data how one affects the other.

P T.O

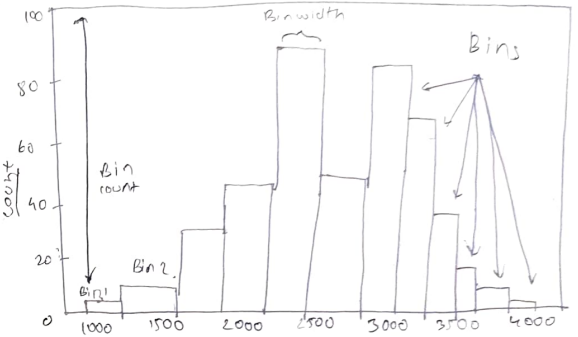


Scatter plot.

ii) Histogram

- In histogram, the data are grouped into ranges (eg. (0-19, 20-29) & then plotted as connected ~~graphs~~ bars.
- Each bar represents a range of data.
- The width of each bar is proportional to the width of each category & height is proportional to the frequency.

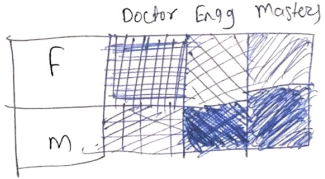
- It provides the visual representation of numerical data by showing the number of data points that falls in specific range called as bins
- Histogram can display a large amount of data and the frequency of the data values
- matplotlib provides a dedicated f^ to compute & display histogram : plt.hist()



Daily steps taken.

Heat map:

- The heatmap visualization uses colour & intensity of the color to show relationship b/w columns.
- Ex→



8] Explain two different data visualization tool function from seaborn & matplotlib

→

① import seaborn as sns.
import matplotbil.pyplot as plt.

```
def bar_plot(data, x, y)
    sns.barplot(data=data, x=x, y=y)
    plt.title("Bar plot")
    plt.show()
```

② import seaborn as sns
import matplotlib.pyplot as plt

```
age = [10, 12, 14, 16, 18, 20]
height = [120, 130, 145, 143, 150, 173]

sns.scatterplot(x=age, y=height)
plt.xlabel("age")
plt.ylabel("height")
plt.show()
```

the frequency

9] write short note on candela, d3.js, Google chart

→ Candela

- Candela is a python-based visualization library built on top of D3.js & plotly, among others.
- It is designed for creating interactive, web-based visualizations easily within Jupyter notebook
- Candela tool works on Kitware's resonant platform that provides many element for data visualization
- Candela is library that provide reusable visualization component for the web.

1] Reusable → Candela is General API not tied to any particular framework or library, so components user create with it can be used from application to application. very easily

2] Visualization → The General api does not provide many constraints. The user must implement a function called render()

3] Component → Use object-oriented concept to implement the notation of visualization component

4] for the web → The output interactive visualization using web technologies such as HTML & JS.

D3.js

- D3.js stands for ~~docum~~ data-driven document. It is javascript library used for creating interactivity dynamic & data-driven visualizations on the web.
- D3.js is quite popular now a days with over 200 million downloaders which provides many visuals such as hierarchical bunding, Treemap etc.....
- D3.js is also useful when ~~data~~ user wants animated & interactive visuals.
- D3 allows to reuse the code which make the work easy for developer.

features

1] Data Binding → D3 binds data to DOM element. ✓

2] Customization → D3 provides full control over the appearance of visual element. Developer can customize aspects such as color, size, ~~sh~~ shape & position

3] Interactivity → Provides ~~interactivity~~ interactive charts with feature like hover effect, zooming, clickable elements.

4] Animation → D3 enables smooth animations & transitions

Google Charts API

- The Google chart API is a Javascript-based tool provided by Google for creating interactive charts & graphs on ~~the~~ web pages.
- It allows user to create simple chart using API. without advanced coding or design skills.

1) Easy to use
2) Interactive
3) Customizable.
4) No backend required.

Q] Give Different Analytical Techniques used in Big Data Visualization.

→ 1] Classification → You use Logistic regression, decision tree & Naive Bayes classifier to classify your data points & then appropriately visualize them. linear topology, Graph topology & tree topology are some of the most suitable formats of creating visualization based on classification.

2] Clustering → Groups.

3] Regression → Relationship between two or more correlated variables. To predict value of one variable when other is given already.

4] Association → To learn relationship about other object.

Q] Explain different tools for data visualization.

→ Candela & D3.js already learned backside.

1] Tableau

→
- Tableau is the platform that provides the visuals of the data. It helps in visualizing useful insights/information from the dataset.
- It is the fastest evolving Buisness Intelligence tool which provides analytical visualization.
- It helps to visvalize data in charts, graphs, & diagram.
- Why use Tableau ?
  - Easy to operate
  - No skilled workers needed
  - Simple installations
  - Inbuild methods to visualize.
  - Can use many different sources of data.
- Tabelau is capable to run on any ~~platform~~ computer.
- Tableau allows user to directly connect databases & data warehouses. The dasboard to share it live on webpage & mobile devices.