KARNATAK LAW SOCIETY'S

GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI – 590008

(An Autonomous Institution under Visvesvaraya Technological University, Belagavi)

(Approved By AICTE, New Delhi)

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

COURSE PROJECT: DATA VISUALISATION

Guided by: Prof. Shrivatsa Perur

Hemanth Tubachi	2GI18IS015
Laxmi Nyamagoud	2GI18IS020
Rachana Kampli	2GI18IS032
Rohan Kokatanur	2GI18IS066

2020-2021

CERTIFICATE

This is to certify that **Mr. Hemanth I T, Ms. Laxmi Nyamagoud, Ms. Rachana Kampli, Mr. Rohan Kokatanur** of **Sixth Semester** bearing **USN: 2GI18IS015, 2GI18IS020, 2GI18IS032, 2GI18IS066** has satisfactorily completed the course in Course activity of Distributed Computing System. It can be considered as a bonafide work carried out for partial fulfillment of the academic requirement of 6th Semester B.E. (Information Science & Engineering) prescribed by KLS Gogte Institute of Technology, Belagavi during the academic year 2020-21.

The report has been approved as it satisfies the academic requirements prescribed for the said degree.

Signature of The Faculty Member **Signature of The HOD.**

Date: 20/06/2021

Course project report and ppt content

1. Title of the seminar
2. Certificate sheet
3. Abstract
4. Content sheet, table of figures
5. Introduction to topic
6. Survey
7. Applications
8. Conclusion
9. References

Marks allocation:

	Batch No. :					
1.	Project Title:	Marks Range	USN			
2.	Problem statement (PO2)	0-1				
3.	Need Analysis, Variables involved (PO1,PO2)	0-2				
4.	Alternate solutions to solve the problem(PO3)	0-3				
5.	Comparison between the solutions and reason for selecting the final solution(PO1,PO3,PO4)	0-4				
6.	Working model of the final solution (PO3,PO12)	0-5				
7.	Report and Oral presentation skill (PO9,PO10)	0-5				
	Total	20				

*** 20 marks is converted to 10 marks for CGPA calculation**

1.Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.

2.Problem Analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences and Engineering sciences.

3.Design/Development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4.Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5.Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6.The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7.Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8.Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9.Individual and team work: Function effectively as an individual and as a member or leader in diverse teams, and in multidisciplinary settings.

10.Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of the engineering management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-long learning: Recognize the need for and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

TITLE OF THE SEMINAR:

Data Visualization

ABSTRACT:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users.

INTRODUCTION

Why we need Data Visualization:

It can provide you some great help in:

- ○ Interpreting data better and memorable.
- ○ Noticing correlations
- ○ Figuring outliers
- ○ Feature Engineering
- ○ Cause-Effect relations

Data Types:

- Categorical variables are the ones that don’t have any ordering e.g. Gender, Grades, Marital Status, Job Position, etc.
- Numerical Variables are segmented into Ordinal and Quantitative variables.
- Ordinal variables are categories that can be ranked. E.g. Satisfaction (Good, Bad, and Average), Potential (High, Medium, and Low), etc.

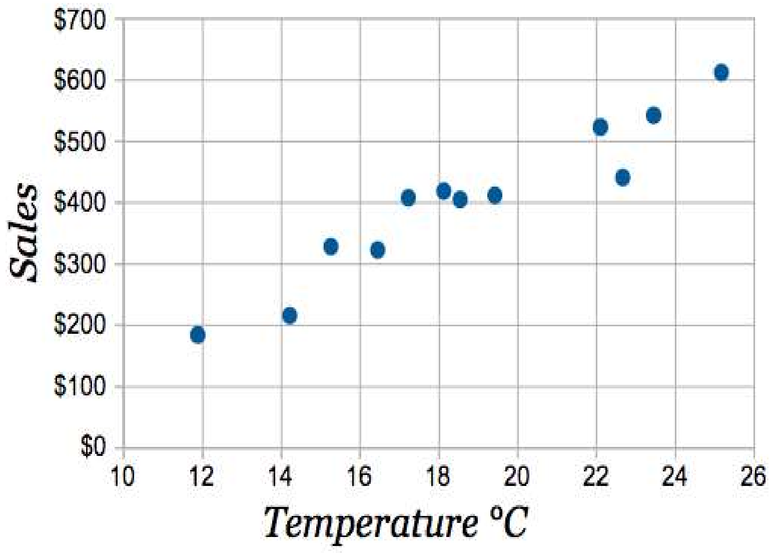
- Quantitative variables are the ones that can take any range of numeric values between -infinity to +infinity. E.g. Age, Salary, Revenue, Sales, etc

Types of Data Visualization:

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards

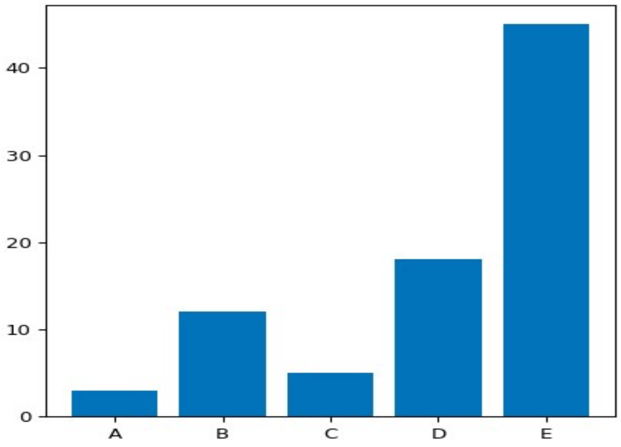
Different types of graphs are:

1. Scatter Plot:



It is basically an X, Y coordinate plots i.e. between two numerical data columns which can be helpful to track down the regression line.

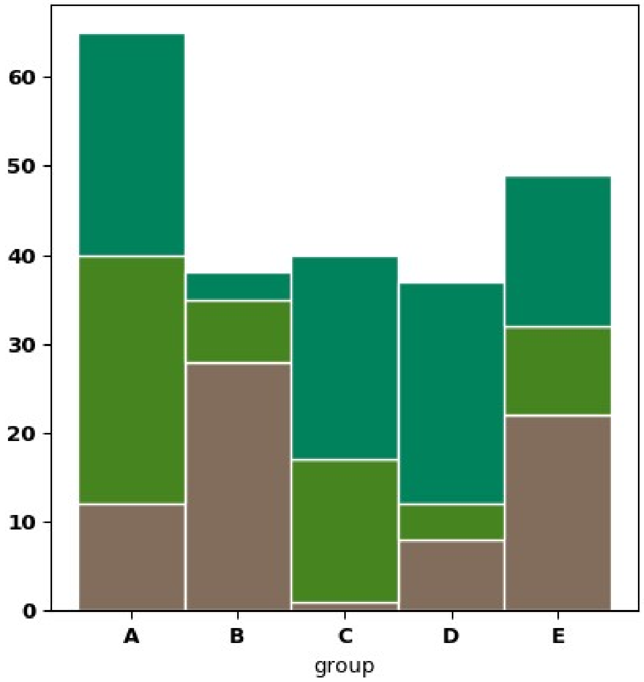
2.Bar Plots:



It is amongst the most popular plots we often encounter. It is used to compare numerical data over some categories/groups.

Example: If we need to compare the number of students passed in different subjects, we might need a barplot. In the above image, y-axis can be taken as Marks and x-axis can be considered as Subjects (A, B, C, D, E).

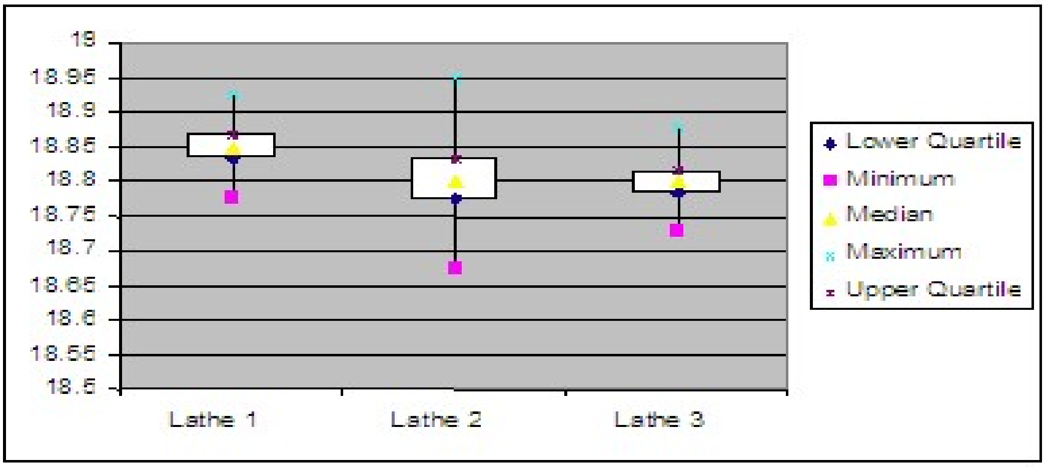
3.Stack Bar Graph:



It might be the case that you want to compare Marks for different sections(suppose A, B, C) and subjects, you have stacked bar plots. Here, we can plot numerical data against groups and subgroups. For us, groups: sections & subgroups: subjects or Vice versa.

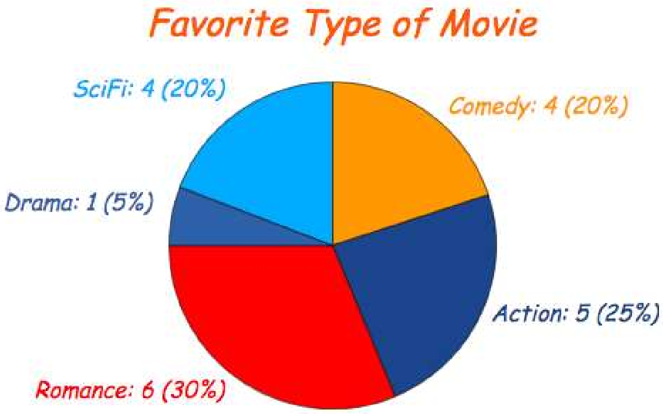
4.Box Plot:

Box plots provide a lot of information about any numerical data column. Its main purpose is to give an idea/summary of the distribution of the data.



5.Pie Chart:

It is again to compare numerical data against a category just like a bar plot but with a difference. It helps us to compare data as a fraction of the whole (percentages rather than raw numbers). In our example, it can be used when we need to find the percentage of students passed in certain subjects and not only numbers.



6.Histogram:

It refers to a graphical representation, that display data by way of bars to show the frequency of numerical data. It indicates distribution of non-discrete variables and presents quantitative data.

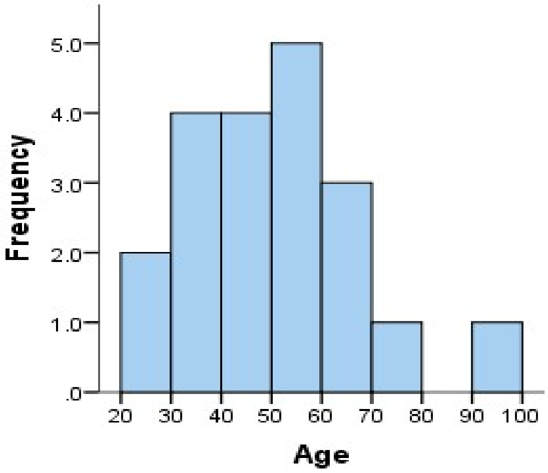


Figure to Illustrates how different graphs can be used to visualize pattern in the data taking into consideration the data type of the variable.

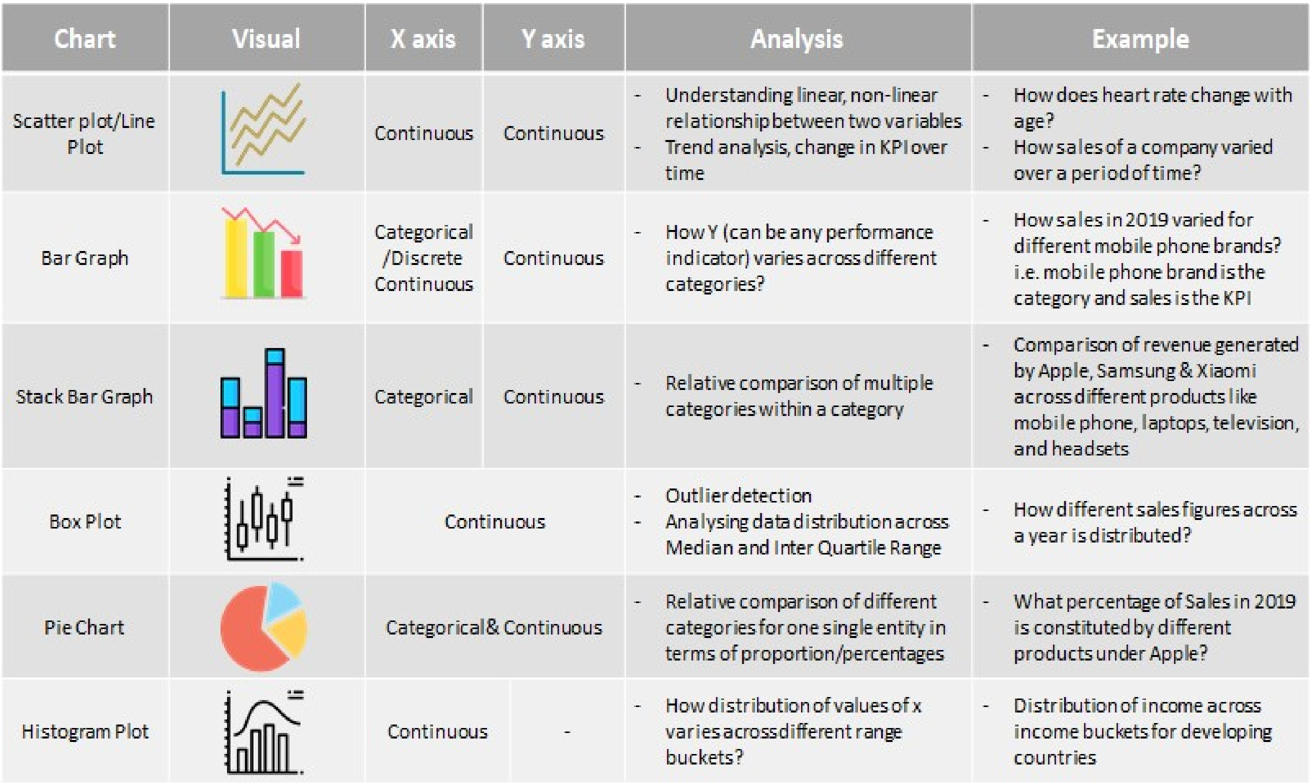


Chart	Visual	X axis	Y axis	Analysis	Example
Scatter plot/Line Plot		Continuous	Continuous	- Understanding linear, non-linear relationship between two variables - Trend analysis, change in KPI over time	- How does heart rate change with age? - How sales of a company varied over a period of time?
Bar Graph		Categorical /Discrete Continuous	Continuous	- How Y (can be any performance indicator) varies across different categories?	- How sales in 2019 varied for different mobile phone brands? i.e. mobile phone brand is the category and sales is the KPI
Stack Bar Graph		Categorical	Continuous	- Relative comparison of multiple categories within a category	- Comparison of revenue generated by Apple, Samsung & Xiaomi across different products like mobile phone, laptops, television, and headsets
Box Plot		Continuous		- Outlier detection - Analysing data distribution across Median and Inter Quartile Range	- How different sales figures across a year is distributed?
Pie Chart		Categorical& Continuous		- Relative comparison of different categories for one single entity in terms of proportion/percentages	- What percentage of Sales in 2019 is constituted by different products under Apple?
Histogram Plot		Continuous	-	- How distribution of values of x varies across different range buckets?	- Distribution of income across income buckets for developing countries

For General:

Why data visualization is important for any career?

It’s hard to think of a professional industry that doesn’t benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

For Data Science:

- Data visualization is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying trends and clusters, spotting local patterns, evaluating modeling output, and presenting results.
- It is essential for exploratory data analysis and data mining to check data quality and to help analysts become familiar with the structure and features of the data before them.

Data Checking And Cleaning:

Data visualization can be used to look for obvious errors in the dataset including nulls, random values, distinct records, the format of dates, sensibility of spatial data, and string and character encoding.

Data Distribution:

Data visualization can be used to understand the distribution of the data, look for central tendencies (mean, median, and mode), understand the presence of outliers using a boxplot, check for skewness, and even understand the impact of winsorization on data distribution.

Model Assumptions:

Linear regression and other classification models follow certain underlying assumptions like data has to be normally distributed, the correlation between different independent variables shouldn’t exist, homoscedasticity of error terms, and many more. Hence visualizations are a key to validating some of these assumptions as well.

Human-in-the-Loop Analytics:

Data scientists often use humans in the loop analytics to get a look and feel of the data, make a hypothesis, run appropriate analytics to validate the hypothesis, and repeat the process till conclusive evidence is determined. E.g. in Python a very popular package Seaborn has a function called pair plot. Pair plots are very useful in determining the relationship between dependent and independent variables. The idea of the visualization is to get a better understanding of the directional sense of if some of the independent variables impact the model results or not.

Dimension Reduction:

While working with multiple variables it is difficult to visualize the data in an n-dimensional space. E.g. in a data set that has different customer attributes (say numerical) it is difficult to plot the customers considering all attributes. In scenarios like this, dimension reduction techniques like Principal Component Analysis (PCA) or Factor Analysis can be useful to bring down the attributes to fewer dimensions. PCA finds linear combinations of variables that best explain the observations whereas Factor analysis finds linear combinations of variables that best explain the relationship between the variables. The reduced dimension can then be plotted to analyze the customers in a 2D space.

CONCLUSION:

Data visualization forms the backbone of all analytical projects. It not only helps in gaining insights into the data but can be used as a tool for data pre-processing. Having the right set of visualizations for different data types and business scenarios is the keto effective communication of results.

REFERENCES:

- https://towardsdatascience.com/data-visualization-in-data-science-5681cbdde5bf
- https://medium.com/data-science-in-your-pocket/data-visualization-for-data-science-beginners-84bacdb8d72e