KIIT POLYTECHNIC

LECTURE NOTES

ON

Data Science & Analytics

6th Semester, Computer Science & Engineering

Prepared by

Sunil Kumar Sahoo

Lecturer, Department of Computer Science & Engineering

KIIT Polytechnic, Bhubaneswar

Email Id-sunilfcs@kp.kiit.ac.in

CONTENTS

UNIT-I
INTRODUCTION TO DATA SCIENCE

Q1. What is Data Science? Explain different terminologies used in data science.

Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data.

Algorithms

An algorithm is a set of instructions we give a computer so it can take values and manipulate them into a usable form. An algorithm is a set of instructions designed to perform a specific task.

Big Data

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), rate of growth (velocity) and consistency (veracity) and value make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools.

Machine Learning

A process where a computer uses an algorithm to gain understanding about a set of data, then makes predictions based on its understanding. There are many types of machine learning techniques; most are classified as either supervised or unsupervised techniques.

Classification

Classification is a supervised machine learning problem. It deals with categorizing a data point based on its similarity to other data points. You take a set of data where every item already has a category and look at common traits between each item. You then use those common traits as a guide for what category the new item might have.

Database

As simply as possible, this is a storage space for data. We mostly use databases with a Database Management System (DBMS), like SQL or MySQL. These are computer applications that allow us to interact with a database to collect and analyze the information inside.

Data Warehouse

A data warehouse is a system used to do quick analysis of business trends using data from many sources. They're designed to make it easy for people to answer important statistical questions without a Ph.D. in database architecture.

Data Wrangling

The process of conversion of data, often through the use of scripting languages, to make it easier to work with is known as data Wrangling or data munging.

Web Analytics

Statistical or machine learning methods applied to web data such as page views, hits, clicks, and conversions (sales), generally with a view to learning what web presentations are most effective in achieving the organizational goal (usually sales). This goal might be to sell products and services on a site, to serve and sell advertising space, to purchase advertising on other sites or to collect contact information. Key challenges in web analytics are the volume and constant flow of data. And the navigational complexity and sometimes lengthy gaps that precede users' relevant web decisions.

Artificial Intelligence (AI)

A discipline involving research and development of machines that are aware of their surroundings. Most work in A.I. centers on using machine awareness to solve problems or accomplish some task. In case you didn't know, A.I. is already here: think self-driving cars, robot surgeons, and the bad guys in your favorite video game.

Business Intelligence (BI)

Similar to data analysis, but more narrowly focused on business metrics. The technical side of BI involves learning how to effectively use software to generate reports and find important trends. It's descriptive, rather than predictive. It is a set of methodologies, process, theories that transform raw data into useful information to help companies make better decisions.

Data Analytics

Analytics is the systematic computational analysis of data or statistics. It is used for the discovery, interpretation, and communication of meaningful patterns in data.

2. Discuss various types of data science toolkit in detail.

A Data Scientist is responsible for extracting, manipulating, pre-processing and generating predictions out of data. In order to do so, he requires various statistical tools and programming languages.

Top Data Science Tools

1-**R**

R is a programming language used for data manipulation and graphics. Originating in 1995, this is a popular tool used among data scientists and analysts. It is the open source version of the S language widely used for research in statistics. According to data scientists, R is one of the easier languages to learn as there are numerous packages and guides available for users.

2-Python

Python is another widely used language among data scientists, created by Dutch programmer Guido Van Rossum. It's a general-purpose programming language, focusing on readability and simplicity. If you are not a programmer but are looking to learn, this is a great language to start with. It's easier than other general-purpose languages.

3-Keras

Keras is a deep learning library written in Python. It runs on TensorFlow allowing for fast experimentation. Keras was developed to make deep learning models easier and helping users treat their data intelligently in an efficient manner.

4. SAS (STATISTICAL ANALYSIS SYSTEM)

It is one of those data science tools which are specifically designed for statistical operations. **SAS is a closed source proprietary software** that is used by large organizations to analyze data. SAS uses base SAS programming language which for performing statistical modeling. It is widely used by professionals and companies working on reliable commercial software

5. Apache Spark

Apache Spark is general purpose cluster computing system. It provides high-level API in Java, Scala, Python, and R. Spark provides an optimized engine that supports general execution graph. It also has abundant high-level tools for structured data processing, machine learning, graph processing and streaming. The Spark can either run alone or on an existing cluster manager.

3. BigML

BigML, it is another widely used Data Science Tool. It provides a fully intractable, cloud-based GUI environment that you can use for processing Machine Learning Algorithms**.** BigML provides a standardized software using cloud computing for industry requirements. Through it, companies can use Machine Learning algorithms across various parts of their company.

4. D3

Javascript is mainly used as a client-side scripting language**.** D3.js, a Javascript library allows you to make interactive visualizations on your web-browser. With several APIs of D3.js, you can use several functions to create dynamic visualization and analysis of data in your browser.

5. MATLAB

MATLAB is a multi-paradigm numerical computing environment for processing mathematical information. It is a closed-source software that facilitates matrix functions, algorithmic implementation and statistical modeling of data. MATLAB is most widely used in several scientific disciplines.In Data Science, MATLAB is used for simulating **neural networks** and fuzzy logic. Using the MATLAB graphics library, you can create powerful visualizations. MATLAB is also used in image and signal processing. This makes it a very versatile tool for Data Scientists as they can tackle all the problems, from data cleaning and analysis to more advanced Deep Learning algorithms. Furthermore, MATLAB's easy integration for enterprise applications and embedded systems make it an ideal Data Science tool. It also helps in automating various tasks ranging from extraction of data to re-use of scripts for decision making. However, it suffers from the limitation of being a closed-source proprietary software.

6. Jupyter

Project **Jupyter** is an open-source tool based on IPython for helping developers in making open-source software and experiences interactive computing. Jupyter supports multiple languages like Julia, **Python**, and R. It is a web-application tool used for writing live code, visualizations, and presentations. Jupyter is a widely popular tool that is designed to address the requirements of Data Science.

7. Matplotlib

Matplotlib is a plotting and visualization library developed for Python. It is the most popular tool for generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc.

8. NLTK

Natural Language Processing has emerged as the most popular field in Data Science. It deals with the development of statistical models that help computers understand human language. These statistical models are part of Machine Learning and through several of its algorithms, are able to assist computers in understanding natural language. NLTK is widely used for various language processing techniques like tokenization, stemming, tagging, parsing and machine learning.

9. Scikit-learn

Scikit-learn is a library based in Python that is used for implementing Machine Learning Algorithms. It is simple and easy to implement a tool that is widely used for analysis and data science. It supports a variety of features in Machine Learning such as data preprocessing, classification, regression, clustering, dimensionality reduction, etc

10. TensorFlow

TensorFlow has become a standard tool for Machine Learning. It is widely used for advanced machine learning algorithms like Deep Learning. Developers named TensorFlow after Tensors which are multidimensional arrays. It is an open-source and ever-evolving toolkit which is known for its performance and high computational abilities. TensorFlow can run on both CPUs and GPUs and has recently emerged on more powerful TPU platforms.

Q. Explain the application of Data Science in detail.

Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data.

- Airline Route Planning
- Fraud and Risk Detection
- Healthcare
- Internet Search
- Targeted Advertising
- Website Recommendations
- Advanced Image Recognition
- Speech Recognition
- Gaming

Airline Route Planning

Airline Industry across the world is known to bear heavy losses. Except for a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air-fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements. Now using data science, the airline companies can:

- Predict flight delay.
- Decide which class of airplanes to buy.
- Whether to directly land at the destination or take a halt in between (For example, A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.
- Effectively drive customer loyalty programs
- Southwest Airlines, Alaska Airlines are among the top companies who've embraced data science to bring changes in their way of working

Fraud and Risk Detection

The earliest applications of data science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the

initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

Healthcare

The healthcare sector, especially, receives great benefits from data science applications.

Internet Search

Now, this is probably the first thing that strikes your mind when you think Data Science Applications. When we speak of search, we think 'Google'. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL, and so on. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in a fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data every day.

Targeted Advertising

If you thought Search would have been the biggest of all data science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms. This is the reason why digital ads have been able to get a lot higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user's past behavior. This is the reason why you might see ads of Data Science Training Programs while I see an ad of apparels in the same place at the same time.

Website Recommendations

Aren't we all used to the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them but also adds a lot to the user experience. A lot of companies have fervidly used this engine to promote their products in accordance with user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, Linkedin, imdb and many more use this system to improve the user experience. The recommendations are made based on previous search results for a user.

Advanced Image Recognition

You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm.
In their latest update, Facebook has outlined the additional progress they've made in this area, making specific note of their advances in image recognition accuracy and capacity.
In addition, Google provides you with the option to search for images by uploading them. It uses image recognition and provides related search results.

Speech Recognition

Some of the best examples of speech recognition products are Google Voice, Siri, Cortana etc. Using speech-recognition feature, even if you aren't in a position to type a message, your life wouldn't stop. Simply speak out the message and it will be converted to text. However, at times, you would realize, speech recognition doesn't perform accurately.

.

Gaming

Games are now designed using machine learning algorithms which improve/upgrade themselves as the player moves up to a higher level. In motion gaming also, your opponent (computer) analyzes your previous moves and accordingly shapes up its game. EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using data science.

Q. Explain the different data types in Data Science in detail.

TYPES OF DATA

Thus Data and Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of four types.

1. Unstructured data: Word, PDF, Text, images, audio and video
2. Semi Structured data: XML data.
3. Meta Data: Data about data
4. Structured data: Relational data.

Unstructured Big Data

-

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, audio and videos etc.

Semi-Structured data

Semi structured data can contain both the forms of data. We can see Semi structured data in form but it is actually not defined .With example a table definition in relational DBMS.
Example of semi-structured data is a data represented in XML file. Web pages are generated in scripting of HTML which is also an example semi structured data.

Personal data stored in a XML file
<rec><name>Amitav</name><gender>Male<gender><age>45</age></rec>
<rec><name>Sudipta</name><gender>Male</gender><age>17</age></rec>
<rec><name>Soumya</name><gender>Male</gender><age>15</age></rec>

Meta Data

Metadata is defined as the data providing information about one or more aspects of the data. It is used to summarize basic information about data which can make tracking and working with specific data easier.

There are three main types of metadata:

- **Descriptive metadata** describes a resource identification It can include elements such as title of the book, abstract and keywords.

- **Structural metadata** indicates how compound objects are put together e.g. how pages are ordered to form chapters.

- **Administrative metadata** provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

Structured Data

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'Structured' data. In other words all data which can be stored in database SQL in form of table with rows and columns.

An employee table is an example of structured data

Employee_Id	Employee_name	Gender	Dept	Salary
K001	SUJIT	M	FINANCE	50,000
K002	DIPTA	M	ADMIN	60,000
K003	SOUMYA	F	FINANCE	55,000

UNIT-II

DATA MANAGEMENT PLAN

A data management plan describes how research data are collected or created, how data are used and stored during research and how made accessible for others after the research has been completed.

1. Describe briefly what kind of data will be collected and how they will be collected.

2. Outline the type(s) of data (e.g. survey, interview, observation, face-to-face focus group, self-administered writings/diaries, photographs, news articles etc.) and estimate the foreseeable amount/volume of each data type.

3. Describe also any existing data you will reuse.

Data Collection Methods

The choice of data collection methods depends on the research problem under study and the information gathered about the variable. Broadly, the data Collection Methods classified into two categories:

1. Primary Data Collection Methods: The primary data are the first by the researcher for the first time and is original in nature. The researcher collects the fresh data when the research problem is unique, and no related research work is done by any other person. The results of the research are more accurate when the data is collected directly by the researcher but however it is costly and time-consuming.

2. Secondary Data Collection Methods: When the data is collected by someone else for his research work and has already passed through the statistical analysis is called the secondary data. Thus, the secondary data is the second-hand data which is readily available from the other sources. One of the advantages of using the secondary data is that it is less expensive and at the same time easily valuable, but however the authenticity of the findings can be questioned. Thus, the researcher can obtain data from either of the sources depending on the nature of his study and the pursued research objective.

APPLICATION PROGRAMMING INTERFACE (API)

An API is a set of subroutine definitions, protocols, and tools for building application software. In general terms, it is a set of clearly defined methods of communication between various software components. A good API makes it easier to develop a computer program by all the building blocks. Which are then put together by the programmer. An API may be for a web based system, operating system, database system, and computer hardware or software library. An API specification can take many forms, but often includes specifications for Structures object classes, variables or remote calls. POSIX, Window ASPI.

The Modern API

1. Modern APIS adhere to standards that are easily accessible and understood broadly.

2. They are treated more like products than code. They are designed for consumption for specific audiences (e.g., mobile developers), they are documented, and they are versioned in a way that users can have certain expectations of its maintenance and lifecycle.

3. Because they are much more standardized, they have a much stronger discipline for Security and governance,

4 Modern APIs are well documented for consumption and versioning

STORAGE MANAGEMENT

The term storage management encompasses the technologies and processes organization to maximize or improve the performance of their data storage resources. It is a broadly categories that includes virtualization, replication, mirroring, security, compression, traffic analysis automation, storage provisioning and related techniques. By some estimates, the amount of digital information stored in the world's computer system is doubling every year. As a result, organizations feel constant pressure to expand their capacity. However, doubling a company's storage capacity every year is an expensive) In order to reduce some of those costs and improve the capabilities and security of solutions, organizations turn to a variety of storage management solutions.

Storage Management Benefits

Many storage management technologies, like storage virtualization, de-duplication and compression allow companies to better utilize their existing storage. The benefits of these approaches lower costs both the one-time capital expenses associated with storage devices and the operational costs for maintaining those devices. Most storage management techniques also simplify the management of storage devices. That can allow companies to save time and even reduce the number of IT workers need to maintain their storage systems, which in turn, also reduces overall storage .

Storage Resource Management (SRM)

Storage management is very closely related to Storage Resource Management (SRM). SRM often refers particularly to software used to manage storage networks and devices. By contrast, the term "storage management" can refer to devices and processes, as well as actual software. In addition, SRM usually refers specifically to software for allocating storage capacity based on company policies and ongoing events. It may include asset management, charge back, and capacity management, configuration management, data and media migration, event management. Performance and availability management, policy management, quota management and media management capabilities. In short, SRM is a subset of storage management; however, the two terms are sometimes used interchangeably.

Mass Storage Management

Mass storage refers to various techniques and devices for storing large amounts of data. The earliest storage devices were punched paper cards, which were used as early as 1804 to control silk-weaving looms. Modern mass storage devices include all types of disk drives and tape drives. Mass storage is distinct from memory, which refers to temporary storage areas within the computer. Unlike main memory. Mass storage devices retain data even when the computer is turned off.

Examples of Mass Storage Devices (MSD)

Common types of mass storage include the following:

1. Solid-state drives (SSD)
2. Hard drives
3. External hard drives
4. Optical drives
5. Tape drives
6. RAID storage
7. USB storage
8. Flash memory cards

UNIT-III

Q) KEY DIFFERENCE BETWEEN DATA ANALYSIS AND DATA ANALYTICS

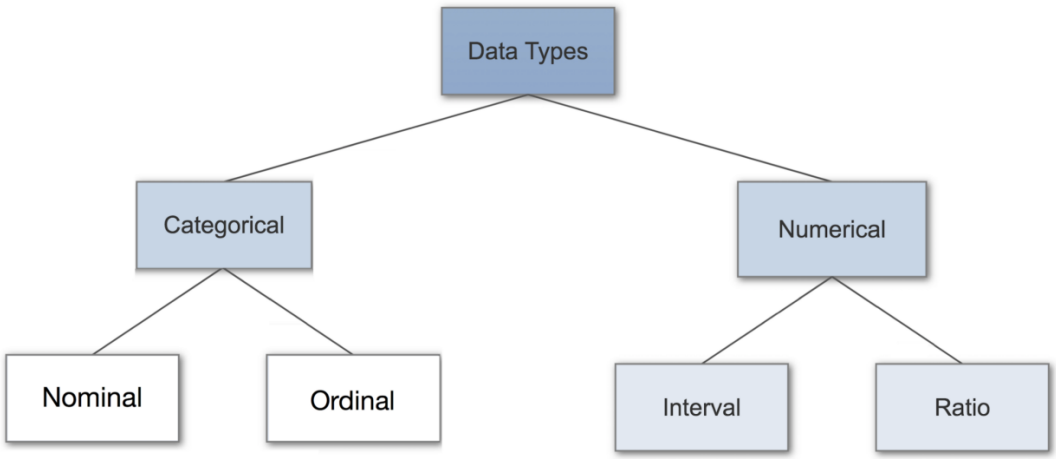
1) Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight.1) Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.

2) Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not, and what was expected from a product or service. 2) Data analytics helps businesses in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies. It helps in business growth by reducing risks, costs, and making the right decisions.

3) In data analysis, experts explore past data, break down the macro elements into the micros with the help of statistical analysis, and draft a conclusion with deeper and significant insights. 3) Data analytics utilizes different variables and creates predictive and productive models to challenge in a competitive marketplace.

Q) What is statistical data? Explain Types of statistical data

Data Types are an important concept of statistics, which needs to be understood, to correctly apply statistical measurements to your data and therefore to correctly conclude certain assumptions about it.



Categorical Data

Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0(zero) for male). Note that those numbers don't have mathematical meaning.

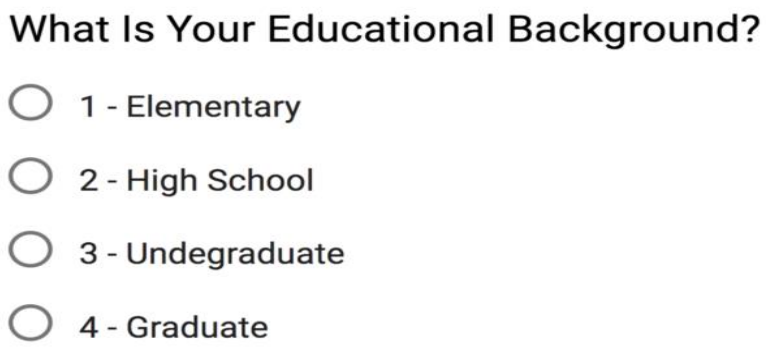
Nominal Data

Nominal values represent discrete units and are used to label variables that have no quantitative value. Just think of them as „labels". Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features below:

The left feature that describes if a person is married would be called „dichotomous", which is a type of nominal scales that contains only two categories.

Ordinal Data

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:



Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

Numerical Data

These data have meaning as a measurement, such as a person 's height, weight, IQ, or blood pressure; or they're a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favorite book before you fall asleep. (Statisticians also call numerical data quantitative data.)

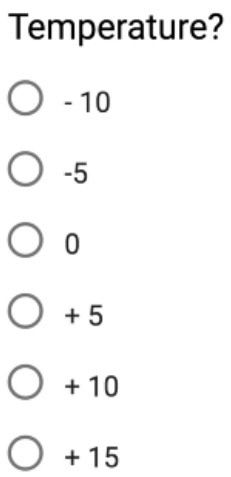
Numerical data can be further broken into two types: discrete and continuous.

Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countable infinite). For example, the number of heads in 100 coin flips takes on values from 0 through 100 (finite case), but the number of flips needed to get 100 heads takes on values from 100 (the fastest scenario) on up to infinity (if you never get to that 100th heads).

Continuous data represent measurements, their possible values cannot be counted and can only be described using intervals on the real number line. For example. the exact amount of gas Purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons represented by the intervals [0, 20].

Interval Data

Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see below:



The problem with interval values data is that they **don't have a „true zero"**. That means in regards to our example, that there is no such thing as no temperature. With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios. Because there is no true zero, a lot of descriptive and inferential statistics can't be applied.

Ratio Data

Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero**. Good examples are height, weight, length etc.

Q) **What is big data analytics? Explain types of big data analytics**

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different form the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways.

Types of Big Data Analytics

a) Descriptive Analytics

It consists of asking the question: What is happening?

It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

b) Diagnostic Analytics

It consists of asking the question: Why did it happen? Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

ç) Predictive Analytics

It consists of asking the question: What is likely to happen?

It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

d) Prescriptive Analytics

It consists of asking the question: What should be done? It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

Unit-IV

Q. Explain the techniques of data visualisation in data science.

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. It is one of the steps in data analysis or data science.

Data visualization techniques

Charts

The easiest way to show the development of one or several data sets is a chart. Charts vary from bar and line charts that show relationship between elements over time to pie charts that demonstrate the components or proportions between the elements of one whole.

Plots

Plots allow to distribute two or more data sets over a 2D or even 3D space to show the relationship between these sets and the parameters on the plot. Plots also vary: scatter and bubble plots are the most traditional. Though when it comes to big data, analysts use box plots that enable to visualize the relationship between large volumes of different data.

Maps

Maps are widely-used in different industries. They allow to position elements on relevant objects and areas - geographical maps, building plans, website layouts, etc. Among the most popular map visualizations are heat maps, dot distribution maps, cartograms.

Diagrams and matrices

Diagrams are usually used to demonstrate complex data relationships and links and include various types of data on one visualization. They can be hierarchical, multidimensional, and tree-like. Matrix is a big data visualization technique that allows to reflect the correlations between multiple constantly updating (steaming) data sets.

Q Write a short notes

A .Data visualization

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 has become the de facto standard for modern business intelligence (BI). The success of the two leading vendors in the BI space, Tableau and Qlik -- both of which heavily emphasize visualization -- has moved other vendors toward a more visual approach in their software. Virtually all BI software has strong data visualization functionality. Data visualization tools have been important in democratizing data and analytics and making data-driven insights available to workers throughout an organization. They are typically easier to operate than traditional statistical analysis software or earlier versions of BI software. This has led to a rise in lines of business implementing data visualization tools on their own, without support from IT.

B. Data Encoding

Encoding is the process of converting the **data** or a given sequence of characters, symbols, alphabets etc., into a specified format, for the secured transmission of **data**. Decoding is the reverse process of **encoding** which is to extract the information from the converted format.

Encoding Techniques

The data encoding technique is divided into the following types, depending upon the type of data conversion.

- **Analog data to Analog signals** − The modulation techniques such as Amplitude Modulation, Frequency Modulation and Phase Modulation of analog signals, fall under this category.

- **Analog data to Digital signals** − This process can be termed as digitization, which is done by Pulse Code Modulation *PCM*. Hence, it is nothing but digital modulation. As we have already discussed, sampling and quantization are the important factors in this. Delta Modulation gives a better output than PCM.
- **Digital data to Analog signals** − The modulation techniques such as Amplitude Shift Keying *ASK*, Frequency Shift Keying *FSK*, Phase Shift Keying *PSK*, etc., fall under this category. These will be discussed in subsequent chapters.
- **Digital data to Digital signals** − These are in this section. There are several ways to map digital data to digital signals.

Q Explain Data encoding strategies in data science with a suitable examples

"In many practical data science activities, the data set will contain categorical variables. These variables are typically stored as text values". (Practical Business Python) Since machine learning is based on mathematical equations, it would cause a problem when we keep categorical variables as is. Many algorithms support categorical values without further manipulation, but in those cases, it's still a topic of discussion on whether to encode the variables or not. The algorithms that do not support categorical values, in that case, are left with encoding methodologies. Let's discuss some methods here.

Encoding Methodologies

Label encoding

In label encoding, we map each category to a number or a label. The labels chosen for the categories have no relationship. So categories that have some ties or are close to each other lose such information after encoding.

Frequency encoding

It is a way to utilize the frequency of the categories as labels. In the cases where the frequency is related somewhat with the target variable, it helps the model to understand and assign the weight in direct and inverse proportion, depending on the nature of the data.

One - hot encoding

In this method, we map each category to a vector that contains 1 and 0 denoting the presence of the feature or not. The number of vectors depends on the categories which we want to keep. For high cardinality features, this method produces a lot of columns that slows down the learning significantly. There is a buzz between one hot encoding and dummy encoding and when to use one. They are much alike except one hot encoding produces the number of columns

equal to the number of categories and dummy producing is one less. This should ultimately be handled by the modeller accordingly in the validation process.

Target or Impact or Likelihood encoding

Target Encoding is similar to label encoding, except here labels are correlated directly with the target. For example, in mean target encoding for each category in the feature label is decided with the mean value of the target variable on a training data. This encoding method brings out the relation between similar categories, but the relations are bounded within the categories and target itself. The advantages of the mean target encoding are that it does not affect the volume of the data and helps in faster learning and the disadvantage is its harder to validate. Regularization is required in the implementation process of this encoding methodology. Visit target encoder in python and R.

What are the conventional data visualization tools?

Best Data Visualization Techniques for small and large data

Bar Chart. Bar charts are used for comparing the quantities of different categories or groups.

Pie and Donut Charts.

Histogram Plot

Scatter Plot.

Visualizing Big Data.

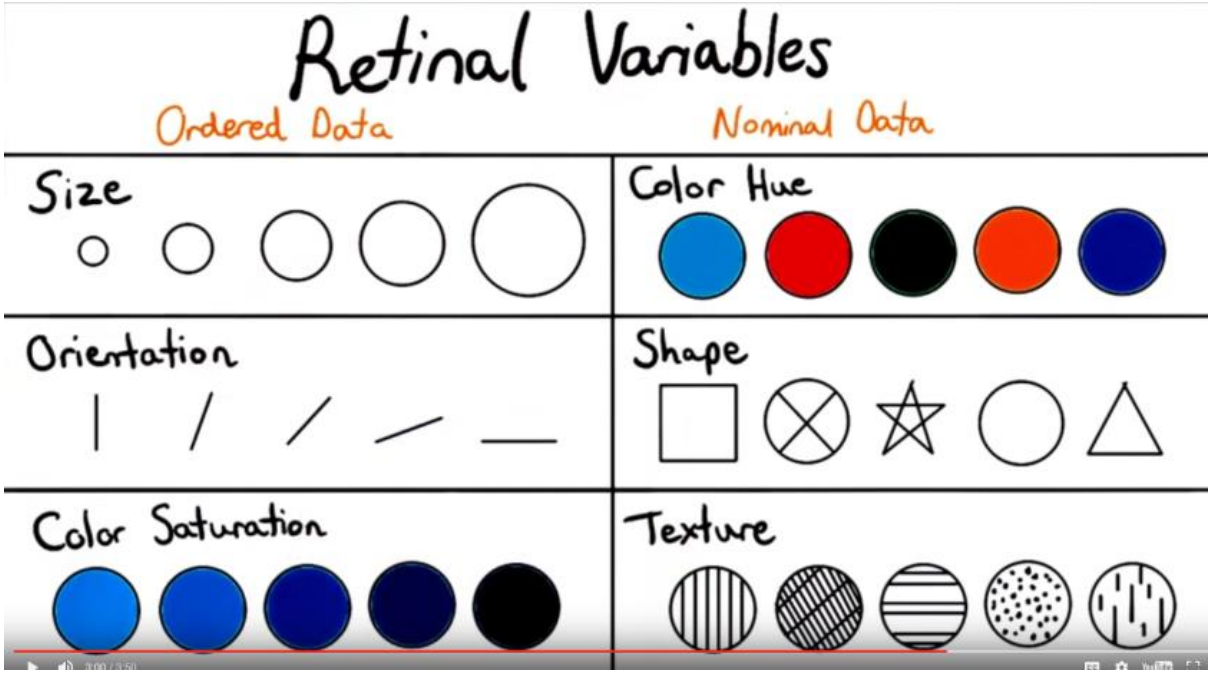
Box and Whisker Plot for Large Data.

Word Clouds and Network Diagrams for Unstructured Data.

Correlation Matrices.

Q. Explain Retinal variables.

The retinal variables are size (length and area), shape, texture, color, orientation (or slope), and value. Each variable can be classified using points, lines and areas. Moreover, color may be described by hue, saturation and brightness, and attributes such as transparency and animation may be added. We can use retinal variable to encode more information to a position element. Size is an example of retinal variable. It is good for ordered nominal data.



Q. Explain Importance of data visualization

1. Identify areas that need attention or improvement
2. Clarify which factors influence customer behavior.
3. Help you understand which products to place where.
4. Predict sales volumes.
5. Data visualization solutions were initially developed as a business tool for enterprises
6. That could afford to hire business analysts, citizen data scientists and BI experts capable.

Graphical displays should:

1. Show the data
2. Induce the viewer to think about the substance rather than about methodology, design, the technology of graphic production or something else
3. Avoid distorting what the data has to say
4. Present many numbers in a small space
5. Make large data sets coherent
6. Encourage the eye to compare different pieces of data.

UNIT-V

What is Python?

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

It is used for:

- web development (server-side)
- software development
- mathematics
- System scripting.

Characteristics of Python

Following are important characteristics of **Python Programming** −

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Applications of Python

As mentioned before, Python is one of the most widely used language over the web. I'm going to list few of them here:

- **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** − Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
- **A broad standard library** − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

- Extendable − You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- Databases − Python provides interfaces to all major commercial databases.
- GUI Programming − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- Scalable − Python provides a better structure and support for large programs than shell scripting.

Built-in Data Types

In programming, data type is an important concept.

Variables can store data of different types, and different types can do different things.

Python has the following data types built-in by default, in these categories:

Numeric Types:	int, float, complex
Sequence Types:	list, tuple, range
Mapping Type:	dictionary
Set Types:	set
Boolean Type:	bool
Text Type:	str

```
a=10
b=2.5
f=float(a)
print(f)
i=int(b)
print(i)
```

List

Lists are used to store multiple items in a single variable.

Lists are one of 4 built-in data types in Python used to store collections of data, the other 3 are Tuple, Set, and Dictionary, all with different qualities and usage.

Lists are created using square brackets:

```
L=["apple","banana","cherry"]
print(L)
```

Example

```
sub=['phy','chem',96,96.5]
print(sub)
print(len(sub))#lenth of the list
print(sub[::-1])#it will print reverse of the list
print(sub*2)#this will repeat the list twice
print(sub[0:2])
print(sub[0:3])
```

String

```
A="welcome to python tutorial"
print(A)
print(len(A))
print(A[8:10])
print(A[::-1])
print(A.lower())
print(A.upper())
```

SET

```
A={1,2,3}
print(A)
B={3,4,5,6}
print(B)
print(A|B)#union
print(A&B)#intersection
```

output

{1, 2, 3}

{3, 4, 5, 6}

{1, 2, 3, 4, 5, 6}

output

{1, 2, 3}

{3, 4, 5, 6}

{3}

Python For Loops

A for loop is used for iterating over a sequence (that is either a list, a tuple, a dictionary, a set, or a string).

This is less like the for keyword in other programming languages, and works more like an iterator method as found in other object-orientated programming languages.

With the for loop we can execute a set of statements, once for each item in a list, tuple, set etc.

```
fruits = ["apple", "banana", "cherry"]
for x in fruits:
 print(x)
```

output

apple

banana

cherry

```
for x in "banana":
 print(x)
```

```
fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
  if x == "banana":
    break
```

output

apple

banana

```
fruits = ["apple", "banana", "cherry"]
for x in fruits:
  if x == "banana":
    break
  print(x)
```

```
fruits = ["apple", "banana", "cherry"]
for x in fruits:
  if x == "banana":
    continue
          print(x)
```

```
for x in range(6):
  print(x)
```

```
for x in range(2, 6):
  print(x)
```

```
for x in range(2, 30, 3):
  print(x)
```

Python while Loop

```
i = 1
while i < 6:
  print(i)
  i += 1
```

```
i = 1
while i < 6:
  print(i)
  if i == 3:
```

```
    break
  i += 1
```

What is a Module?

A module is a file consisting of python code.

A module can define functions, classes and variables.

Example

Save this code in a file named **mymod.py**

```
def greeting(name):
  print("Hello, " + name)
```

Example

Import the module named mymod, and call the greeting function:

```
import mymodule

mymodule.greeting("Jonathan")
```

User Drfined Module

Example

```
def fun(x,y):
  z=x+y
  print(z)


def fun1(sarkar):
  print(sarkar)


import mymod
mymod.fun(10,20)
mymod.fun1("hellow")
```

Built in module

```
import calendar

cal=calendar.month(2019,1)

print(cal)
```

output

```
 January 2019

Mo Tu We Th Fr Sa Su

   1  2  3  4  5  6

 7  8  9 10 11 12 13

14 15 16 17 18 19 20

21 22 23 24 25 26 27

28 29 30 31
```

```
import calendar

cal=calendar.month(2019,1)

x=dir(calendar)

print(cal)

print(x)
```

PYTHON LIBRARY

NumPy (short for Numerical Python) is an open source Python library for doing scientific computing with Python.
It gives an ability to create multidimensional array objects and perform faster mathematical operations. The library contains a long list of useful mathematical functions, including some functions for linear algebra and complex mathematical operations such as Fourier Transform (FT) and random number generator (RNG).

```
import numpy as np

the_array = np.array([49, 7, 44, 27, 13, 35, 71])

an_array = np.where(the_array > 30, 0, the_array)
print(an_array)
```

output- [0 7 0 27 13 0 0]

Pandas is a Python library comprising high-level data structures and tools that has designed to help Python programmers to implement robust data analysis. The utmost purpose of Pandas is to help us identify intelligence in data. Pandas is in practice in a wide range of academic and commercial domains, including finance, neurosciences, economics, statistics, advertising, and web analytic.

```
import pandas as pd

ser1 = pd.Series([1.5, 2.5, 3, 4.5, 5.0, 6])
print(ser1)
```

output

0	1.5
1	2.5
2	3
3	4.5
4	5
5	6

Matplotlib is a versatile Python library that generates plots for data visualization. Matplotlib offers simple and powerful plotting interface, versatile plot types and robust customization. With the diverse plot types and elegant styling options available, it works well for creating professional figures for demonstrations and scientific reports.

```
import matplotlib.pyplot as plt

#Plot a line graph

plt.plot([5, 15])

# Add labels and title

plt.title("Interactive Plot")

plt.xlabel("X-axis")

plt.ylabel("Y-axis")

plt.show()
```

The method bar() creates a bar chart.
The program below creates a bar chart. We feed it the horizontal and vertical (data) data.

```
import numpy as np
import pandas as pd
from pandas import Series, DataFrame
import matplotlib.pyplot as plt
```

```
data = [23, 45, 56, 78, 213]
plt.plot([1,2,3,4,5], data)
plt.show()
```

UNIT-VI

Q.EXPLAIN THE VARIOUS BIGDATA VISUALIZATION TOOLS

1. Google Chart

Google is an obvious benchmark and well known for the user-friendliness offered by its products and Google chart is not an exception. It is one of the easiest tools for visualising huge data sets. Google chart holds a wide range of chart gallery, from a simple line graph to complex hierarchical tree-like structure and you can use any of them that fits your requirement.

2. Tableau

Tableau desktop is an amazing data visualisation tool (SaaS) for manipulating big data and it's available to everyone. It has two other variants "Tableau Server" and cloud-based "Tableau Online" which are dedicatedly designed for big data-related organisations.

3. D3

D3 or Data Driven Document is a Javascript library for big data visualisations in virtually any way you want. This is not a tool, like the others and the user needs a good grasp over javascript to give the collected data a shape. The manipulated data are rendered through HTML, SVG, and CSS, so there is no place for old browsers (IE 7 or 8) as they don't support SVG (Scalable Vector Graphics). D3 is extremely fast and supports large data sets in real-time. It also produces dynamic interaction and animation in both 2D and 3D with minimal overhead. The functional style of D3 allows you to reuse codes through the various collection of components and plug-ins.

4. Fusion Chart

Fusion chart XT is a Javascript charting library for the web and mobile devices, spread across 120 countries with having clients such as Google, Intel, Microsoft and many others. However, you need a bit knowledge on Javascript for implementing it. Technically, it collects data in XML or JSON format and renders it through charts using Javascript (HTML5), SVG and VML format. It provides more than 90 chart styles in both 2D and 3D visual formats with an array of features like scrolling, panning, and animation effects. However, this tool doesn't come for free. Its pricing range starts from $199 (for individual developers or freelancers) for one year and updates with one-month priority support.

5. Highcharts

Highcharts is a charting library written purely in Javascript hence, a bit knowledge of Javascript is necessary for implementing this tool. It uses HTML5, SVG and VML for displaying charts across various browsers and devices like android, iPhone etc. For any execution, it requires two .js files: This tool is efficient enough to process real-time JSON data and represents them as a chart mentioned by the user. If you are an enthusiastic programmer you can download its source code and modify it as per your need.

6. Canvas

Canvas.js is a javascript charting library with a simple API design and comes with a bunch of eye-catching themes. It is a lot faster than the conventional SVG or Flash charts. It also comes with a responsive design so that it can run on various devices like Android, iPhone, Tablets, Windows, Mac etc.

7. Qlikview

Qlik is one of the major players in the data analytics space with their Qlikview tool which is also one of the biggest competitors of Tableau. Qlikview boasts over 40,000 customers spanning across over 100 countries. Qlik is particularly known for its highly customisable setup and a host of features that help create the visualisations much faster. However, the available options could mean there would be a learning curve to get accustomed with the tool so as to use it to its full potential. Apart from its data visualisation prowess, Qlikview also offers analytics, business intelligence and enterprise reporting features. The clean and clutter-free user experience is one of the notable aspects of Qlikview.

8. Datawrapper

Datawrapper is a data visualisation tool that’s gaining popularity fast, especially among media companies which use it for presenting statistics and creating charts. It has an easy to navigate user interface where you can easily upload a CSV file to create maps, charts and visualisations that can be quickly added to reports. Although the tool is primarily aimed at journalists, its flexibility should accommodate a host of applications apart from media usage.

9. Microsoft Power BI

Microsoft Power BI is a suite of business analytics tools from Microsoft primarily meant for analysing data and sharing the insights. It enables you to explore and dig insights out of your data via any device you use – desktops, tablets or smartphones. It helps you derive quick answers from the data and also can connect to on-premises data sources for real time mapping and analysis.

10. Oracle Visual Analyzer

Introduced in 2015, this web-based tool within the Oracle Business Intelligence Cloud Service claimed a spot at the Magic Quadrant Business Intelligence and Analytics Platform report by Gartner. Interactive visuals and highly advanced analysis clubbed with a customisable dashboard are some of the key features of Oracle Visual Analyzer. Being highly scalable, this data visualisation tool is very suitable for enterprises with large-scale deployments where deep insights and well curated reports are essential. Every bit of data carries a story with it and these data visualisation tools are the gateway to fathom the story it tries to tell us. It helps us to understand about the current statistics and the future trends of the market.

Q. Write Short notes on

a) IMPORTANTANCE OF BIG DATA VISUALIZATION.

1. Review Large Amounts Of Data
2. Spot Trends
3. Identify correlations and unexpected relationships
4. Present the data to others

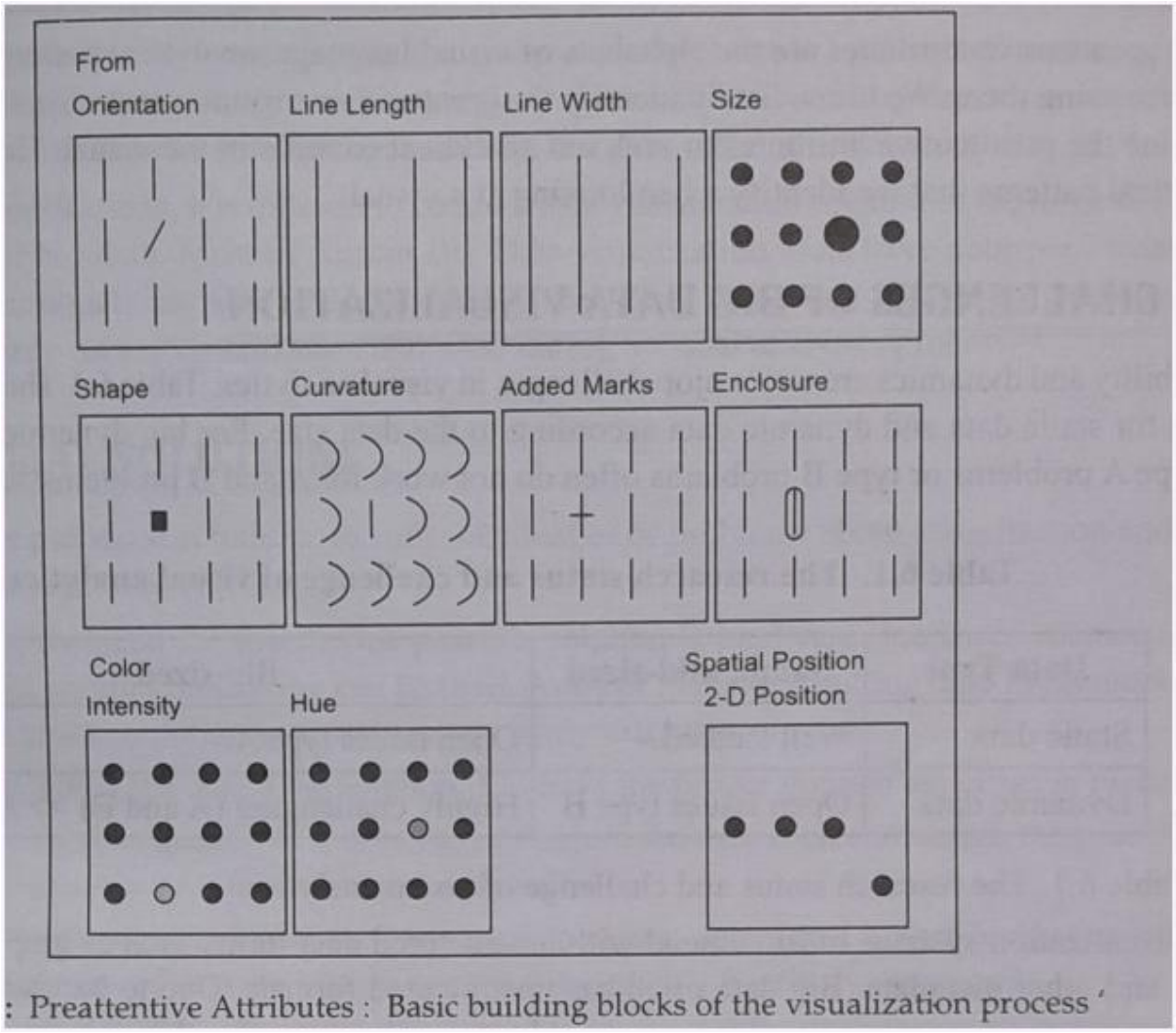
b) KEY ISSUES OF BIG DATA VISUALIZATION

1. Availability of visualization specialists
2. Visualization hardware resources
3. Data quality

c) PREATTENTVE ATTRIBUTES

These attributes are what immediately catch our eye when we look at a visualization. They can be perceived in less than 10 milliseconds, even before we make a conscious effort to notice them.

Here's a list of the preattentive attributes:



: Preattentive Attributes : Basic building blocks of the visualization process '

Considering preattentive attributes can help when deciding which chart type to use for our data. While preattentive attributes are what we immediately identify in a visual, they aren't the only things we notice. We go on to form patterns out of the preattentive attributes.

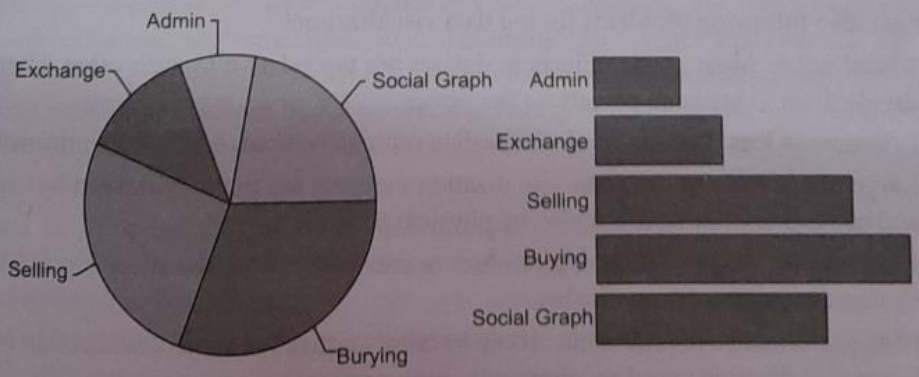


Fig. 6.6: Forming Analytical Patterns Out of Preattentive Attributes

d) CHALLENGES OF BIG DATA VISUALIZATION

There are also following problems for big data visualization:

Visual noise: Most of the objects in dataset are too relative to each other. Users cannot divide them as separate objects on the screen.

Information loss: Reduction of visible data sets can be used, but leads to information loss

Large image perception: Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.

High rate of image change: Users observe data and cannot react to the number of data change or its intensity on display.

High performance requirements: It can be hardly noticed in static visualization because of lower visualization speed requirements and high performance requirement.

e) POTENTIAL SOLUTIONS

Following are the proposed solutions to some challenges or problems about big data visualization

1. Meeting the need for speed: One possible solution is hardware. Increased memory and powerful parallel processing can be used. Another method is putting data in-memory but using a grid computing approach, where many machines are used.

2. Understanding the data: One solution is to have the proper domain expertise in place.

3. Addressing data quality: It is necessary to ensure the data is clean through the process of data governance or information management.

4. Displaying meaningful results: One way is to cluster data into a higher-level view where smaller groups of data are visible and the data can be effectively visualized.

5. Dealing with outliers: Possible solutions are to remove the outliers from the data or create a separate chart for the outliers.

f) FUTURE PROGRESS OF BIG DATA VISUALIZATION

1. **Treemap**: It is based on space filling visualization of hierarchical data.

2. Circle Packing: It is a direct alternative to treemap. Besides the fact that as primitive shape it uses circles, which also can be included into circles from a higher hierarchy level.

3. Sunburst: It uses treemap visualization and is converted to polar coordinate system. The main difference is that the variable parameters are not width and height, but a radius and arc length.

4. Parallel Coordinates: It allows visual analysis to be extended with multiple data factor for different objects.

5. Streamgraph: It is a type of a stacked area graph that is displaced around a central axis resulting in flowing and organic shape.

6. Circular Network Diagram: Data object are placed around a circle and linked by curves based on the rate of their relativeness. The different line width or usually used to measure object relativeness.