# Assignment 3

Ajay Singh
20BAI10300

1. Define data visualization. Illustrate how data visualization is better than traditional text-based data methods.

Sol:

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization can be utilized for a variety of purposes, and it's important to note that is not only reserved for use by data teams. Management also leverages it to convey organizational structure and hierarchy while data analysts and data scientists use it to discover and explain patterns and trends.

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. The practice can also help businesses identify which factors affect customer behavior; pinpoint areas that need to be improved or need more attention; make data more memorable for stakeholders; understand when and where to place specific products; and predict sales volumes.

Data visualization is important because of the processing of information in human brains. Using graphs and charts to visualize a large amount of the complex data sets is more comfortable in comparison to studying the spreadsheet and reports.

Data visualization is an easy and quick way to convey concepts universally. You can experiment with a different outline by making a slight adjustment.

Data visualization is better than traditional text-based data methods because -

- Without data visualization, it is challenging to identify the correlations between the relationship of independent variables.
- It's impossible to make predictions without having the necessary information from the past and present. Trends over time tell us where we were and where we can potentially go.

- Data visualization takes the information from different markets to give you insights into which audiences to focus your attention on and which ones to stay away from.
- Looking at value and risk metrics requires expertise because, without data visualization, we must interpret complicated spreadsheets and numbers. Once information is visualized, we can then pinpoint areas that may or may not require action.
- The ability to obtain information quickly and easily with data displayed clearly on a functional dashboard allows businesses to act and respond to findings swiftly and helps to avoid making mistakes.

2. Comparison between data gathering and data preparation?

Solution:

Data Preparation

Data Preparation is the very first phase of a business intelligence project. It is the phase of transforming raw data into useful information that will later be used for decision-making. Data sources are merged and filtered. They are finally aggregated, and the raw data are subject to the calculation of additional values.

Data Preparation is mainly the phase that precedes the analysis. A graphical user interface that makes the preparation usable is preferably required. Data Preparation is mainly used for an analysis of business data. This involves the collection, cleaning, and consolidation of data. All this takes place in a file that can then be used for the analysis.

This phase is of course essential for filtering unstructured and disordered data. Data Preparation also makes it possible to connect data from different sources, all in real time. Another important advantage of Data Preparation is that it allows you to manage the data collected from a file and to obtain a quick report of this data.

The various data preparation procedures include data collection, which is the initial process for any organization or business. It is at this stage that data is collected from a variety of sources. These sources can really be of any type.

The next step is data discovery. It is then important to understand the data collected in order to classify it into different sets. As the data is often very large, filtering the data can be very time consuming.

It is then equally important to clean and validate the data (data cleansing) in order to remove and discard anything that is not useful for later steps when decision-making is required. Unnecessary or aberrant data should be removed at this stage. Appropriate models should be used to refine the data set. A lock should be used to protect sensitive data.

Once the data has been cleansed, it must go through the test team who will perform all necessary checks. The next step is to define the format of the value entries in order to make the set accessible and understandable to decision-makers. Once all these procedures have been carried out, the data remains to be stored. The analysis tools can then be implemented.

Preparation Data has many advantages. Among other things, it allows a quick response to correct possible errors. The quality of the data is improved, allowing for a more efficient and faster analysis.

Data Gathering

Data gathering is the stage following the preparation phase. The prepared data is then analysed to enable the questions arising from the data preparation to be answered. The data provided is explored interactively. They are reorganized in such a way that they are presented in an understandable way and used by decision-makers. It is therefore a question of exploring data that has not yet been transformed.

Exploration is necessary for decision-makers, who thereby obtain information on data that was previously difficult to perceive. Data mining is in fact the first step in data analysis. It is from this phase that it becomes possible to plan appropriate decisions for the organization or company. This involves identifying and summarizing the main characteristics of a set of data.

A team of experienced analysts is needed to handle visual analysis tools and statistical management software. Sometimes it is necessary to use both manual and automated tools.
Data can be explored manually or automatically. Automated methods are, of course, popular because of their accuracy and speed. Data visualisation tools are particularly effective. Manual data mining allows you to filter and explore data in files such as Excel. Scripting is also used to analyse raw data.
Among the techniques used for Data Exploration is univariate analysis, which is the simplest technique, since only one variable is present in the data. The data is analysed one by one. The analysis here depends on the type of variables, which can be categorical or continuous as the case may be.

3. How can one-dimensional multivariate data be visualized? Explain tree visualization with the help of a suitable example.

Ans

Multivariate analysis is where the fun as well as the complexity begins. Here we analyse multiple data dimensions or attributes (2 or more). Multivariate analysis not only involves just checking out distributions but also potential relationships, patterns, and correlations amongst these attributes. You can also leverage inferential statistics and hypothesis testing, if necessary, based on the problem to be solved at hand to check out statistical significance for different attributes, groups and so on.

The best ways to visualize multivariate data:

- One of the best ways to check out potential relationships or correlations amongst the different data attributes is to leverage a pair-wise correlation matrix and depict it as a heatmap.
- Another way of visualizing multivariate data for multiple attributes together is to use parallel coordinates.
- Another way to visualize the same is to use pair-wise scatter plots amongst attributes of interest. Scatter plots and joint plots are good ways to not only check for patterns, relationships but also see the individual distributions for the attributes.

- Categorical attributes are visualized using separate plots (subplots) or facets for one of the categorical dimensions.
- Another good way is to use stacked bars or multiple bars for the different attributes in a single plot.
- Other plots are histograms or density plots, bar plots, box plot, and violin plots.

Tree Visualization: A decision tree shows a connected hierarchy of boxes to represent the values of records. Records are segmented into groups, which are called nodes. Each node contains records that are statistically like each other with respect to the target field. For example, a node might contain the records for males who have more than 18 years of education. Nodes can then be used to predict a target's field value. For example, the node about males and education might be used to predict salary. Each branch in a decision tree corresponds to a decision rule. To improve performance, due to number of rows in the data source, the analysis is based on a representative sample of the entire data.

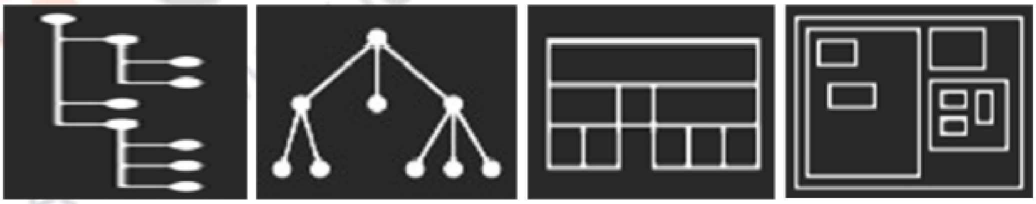

Figure 36.7 Indented lists Node-link trees Layered diagrams Treemaps