**CHAPTER 5**

**DATA VISUALIZATION**

**5.1 Introduction to Data Visualization**

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 sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization is one of the steps of the data mining process, which states that after data has been collected, processed, and modelled, it must be visualized for conclusions to be made. Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format, and deliver data in the most efficient way possible.

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. According to Vitaly Friedman (2008) the "main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key aspects in a more intuitive way. Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualizations which fail to serve their main purpose — to communicate information".

Indeed, Fernanda Vegas and Martin M. Wattenberg suggested that an ideal visualization should not only communicate clearly but stimulate viewer engagement and attention.

Data visualization is closely related to information graphics, information visualization, scientific visualization, exploratory data analysis and statistical graphics. In the new millennium, data visualization has become an active area of research, teaching and development. According to Post et al. (2002), it has united scientific and information visualization.

In the commercial environment data visualization is often referred to as dashboards. Infographics are another very common form of data visualization.

**5.2 Visualizing Graphs**

A Graph is a non-linear data structure consisting of nodes and edges. The nodes are sometimes also referred to as vertices and the edges are lines or arcs that connect any two nodes in the graph.

In this tutorial we are going to visualize undirected Graphs in Python with the help of plotting library.

**5.2.1 Choropleth Maps**

A choropleth map is a type of thematic map in which a set of pre-defined areas is coloured or patterned in proportion to a statistical variable that represents an aggregate summary of a geographic characteristic within each area, such as population density or per-capita income.

Choropleth maps are popular thematic maps used to represent statistical data through various shading patterns or symbols on predetermined geographic areas (i.e. countries). They are good at utilizing data to easily represent variability of the desired measurement, across a region.

A choropleth map of number of developers around the world is shown below:

Map

Description automatically generated

**Figure 5.1 Choropleth Map of Respondents**

**5.2.2 Word Cloud**

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it’s mentioned within a given text and the more important it is.

A word cloud of platforms used by developers is shown below:

Text

Description automatically generated with low confidence

**Figure 5.2 Word Cloud of Platforms**

**5.2.3 Scatter Plot**

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval. For example, weight and height would be on the y-axis, and height would be on the x-axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the dots' pattern from lower left to upper right indicates a positive correlation between the variables being studied. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation.

A 3D Scatter plot of ‘YearsCodePro’, ‘Age’ by their ‘ConvertedComp’ is shown here:

A picture containing chart

Description automatically generated

**Figure 5.3 3D Scatter Plot of ConvertedComp by Age and YearsCodePro**

**5.2.4 Pie Chart**

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

A pie chart of ‘EdLevel’ by percentage of developers is shown as:

Chart, pie chart

Description automatically generated

**Figure 5.4 Pie Chart of Respondents by EduLevel**