**What is Data Visualization?**

In simple terms, data visualization in data science refers to the process of generating graphical representations of information.

These graphical depictions, often known as plots or charts, are pivotal in the realm of data science for effective analysis and interpretation.

Understanding the various types of data visualization in data science is crucial to select the appropriate visual method for the dataset at hand.

Different types serve different analytical needs, from understanding distributions with histograms to spotting trends with line charts.

As one delves deeper into the data science field, the importance of mastering these visualization types becomes even more apparent.

**Why is Data Visualization Important in Data Science?**

Data visualization benefits include communicating your results or findings, monitoring the model’s performance at the evaluation stage, hyperparameter tuning, identifying trends, patterns and correlation between dataset features, data cleaning such as outlier detection, and validating model assumptions.

**Examples of Data Visualization in Data Science**

Here are some popular data visualization examples.

1. **Weather reports:** Maps and other plot types are commonly used in weather reports.

2. **Internet websites:** Social media analytics websites such as Social Blade and Google Analytics use data visualization techniques to analyze and compare the performance of websites.

3. **Astronomy:** NASA uses advanced data visualization techniques in its reports and presentations.

4. **Geography**

5. **Gaming industry**

**What Makes Data Visualization Effective?**

To get the most out of data visualization, you should consider the following things. These are the fundamentals of data visualization.

• **Clarity:** Data should be visualized in a way that everyone can understand. • **Problem domain:** When presenting data, the visualizations should be related to the business problem.

• **Interactivity:** Interactive plots are useful to compare and highlight certain things within the plot.

• **Comparability:** We can compare the thighs easily with good plots. • **Aesthetics:** Quality plots are visually aesthetic.

• **Informative:** A good plot summarizes all relevant information.

**Importance of Data Visualization in Data Science**

Earlier, I mentioned the importance of data visualization in data science. Here are some more details.

**1. Data cleaning**

Data visualization plays an important role in data clearing. Good examples are detecting outliers and removing multicollinearity. We can create scatterplots to detect outliers and generate heatmaps to check multicollinearity.

**2. Data Exploration**

Before building any model, we need to do some exploratory data analysis to identify dataset characteristics. For example, we can create histograms for continuous variables to check for normality in the data. We can create scatterplots between two features to check whether they are correlated. Likewise, we can create a bar chart for the label column with two or more classes to identify class imbalance.

**3. Evaluation of modeling outputs**

We can create a confusion matrix and learning curve to measure the performance of a model during training. Plots are also useful in validating model assumptions. For example, we can create a residuals plot and histogram for the distribution of residuals to validate the assumptions of a linear regression model.

**4. Identifying trends**

Time and seasonal plots are useful in time series analysis to identify certain trends over time.

**5. Presenting results**

As a data scientist, you need to present your findings to the company or other related persons who do not have more knowledge in the subject domain. So, you need to explain everything in plain English. You can use informative plots that summarize your findings. Are you interested in data visualization? Get started with the best Data Science courses.

**Data Visualization Process/Workflow**

The data visualization process or workflow includes the fowling key steps.

**1. Develop your research question**

This may be a business problem or any other related problem that could be solved with a data-driven approach. You should note all the objectives and outcomes plus required resources such as datasets, open-source software libraries, etc.

**2. Get or create your data**

The next step is collecting data. You can use existing datasets if they’re relevant to your research question. Alternatively, you can download **open source datasets** from the internet or do web scraping to collect data.

**3. Clean your data**

Real-world data are messy. So, you need to clean them before using them for visualization. You can identify missing values and outliers and treat them accordingly. You can perform feature selection and remove unnecessary features from the data. You can create a new set of features based on the original features.

**4. Choose a chart type**

The chart type depends on many factors. For example, it depends on the feature type (numerical or categorical). It also depends on the type of visualization you need. Let’s say you have two numerical features. If you want to find their distributions, you can create two histograms for each

feature. If you want to plot their variations, you can create box and whisker plots for each feature. You can create a scatterplot if you want to find a relationship (linear or non-linear, positive or negative) between the two features.

**5. Choose your tool**

You can use open-source data visualization tools such as matplotlib, seaborn, plotty and ggplot. You can also use API-based software such as Matlab, Minitab, SPSS, etc.

**6. Prepare data**

You can extract relevant features. You can do feature standardization if the values of the features are not on the same scale. You can apply data preprocessing steps such as PCA to reduce the dimensionality of the data. That will allow you to visualize high-dimensional data in 2D and 3D plots!

**7. Create a chart**

This is the final step. Here. You define the title and names for the axes. You should also choose a proper chart background to ensure the content is easily readable.

**Tools and Software for Data Visualization**

There are multiple tools and software available for data visualization. **1. Python provides open-source libraries such as**

• Matplotlib

• Seaborn

• Plotty

• Bokeh

• Altair

**2. R provides open-source libraries such as**

• Ggplot2

• Lattice

**3. Other data visualization libraries**

• IBM SPSS

• Minitab

• Matlab for data visualization

• Tableau

• Microsoft Power BI are popular among data scientists.

**Tableau and Microsoft Power BI are popular among data scientists. Data Visualization Techniques in Data Science**

Some of the main data visualization techniques in data science are univariate analysis, bivariate analysis and multivariate analysis.

**1. Univariate Analysis**

In univariate analysis, as the name suggest, we analyze only one variable at a time. In other words, we analyze each variable separately. Bar charts, pie charts, box plots and histograms are common examples of univariate data visualization. Bar charts and pie charts are created for categorical variables, while box plots and histograms are created for numerical variables.

**2. Bivariate Analysis**

In bivariate analysis, we analyze two variables at a time. Often, we see whether there is a relationship between the two variables. The scatter plot is a classic example of bivariate data visualization.

**3. Multivariate Analysis**

In multivariate analysis, we analyze more than two variables simultaneously. The heatmap is a classic example of multivariate data visualization. Other examples are cluster analysis and principal component analysis (PCA).

**Advantages and Disadvantages of Data Visualization Advantages**

There are many advantages of data visualization. Data visualization is used to:

• Communicate your results or findings with your audience • Tune hyperparameters

• Identify trends, patterns and correlations between variables • Monitor the model’s performance

• Clean data

• Validate the model’s assumptions

**Disadvantages**

There are also some disadvantages of data visualization.

• We need to download, install and configure software and open-source libraries. The process will be difficult and time-consuming for beginners. • Some data visualization tools are not available for free. We need to pay for those.

• When we summarize the data, we’ll lose the exact information.

**Essential Skills for Data Visualization**

You should have the following data visualization skills for effective data visualization.

**1. Programming**

You should know R or Python language. R wins, hands down when it comes to data visualization. Its ggplot2 library provides high-level functions to make complex plots with less code. Data visualization in Python can be done using libraries like matplotlib, plotty, bokeh and seaborn for data visualization. Plotty and bokeh can be used for interactive data visualizations.

**2. Software Expertise**

In addition to using R or Python languages, you can also use data visualization software such as Matlab, Minitab and SPSS for data visualization. Data visualization in Excel is also popular. However, they provide limited customizations for your plots. In addition to that, you cannot automate the plot creation process as you can do it with Python or R.

3. Data Science Skills

Data visualization is one of the data science skills. But, for effective data visualization, you need other data science skills such as statistical analysis, data cleaning, processing large data sets, data mining, etc. Data visualization cannot be done alone. It is a collection of these skills.

**4. Public Speaking and Presentation**

When it comes to presenting your findings to the company or other related people, you need to have excellent presentation skills. You should have more confidence when explaining things to a larger audience. For that, you should be familiar with the given problem domain.

**5. Machine Learning**

Machine learning is the ability of computers to learn from data without being explicitly programmed. It is completely different from traditional programming. We can use machine learning algorithms to find important patterns and features in the data. Then, we can visualize those things. There are machine learning algorithms that can be used to perform data cleaning before data visualization. Machine learning is part of the data visualization process.

**Conclusion**

Data visualization is important in every aspect of data visualization in data science. We should clean our data before making any visualization. We should choose the right tool or software that addresses our needs,

such as affordability, ease of use, etc. The main challenge in data visualization is choosing the right plot type. It depends on many factors. Finally, you need excellent public speaking and presentation skills to present your findings.