knowledgehut
upGrad

Visualization in Python

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

Recap

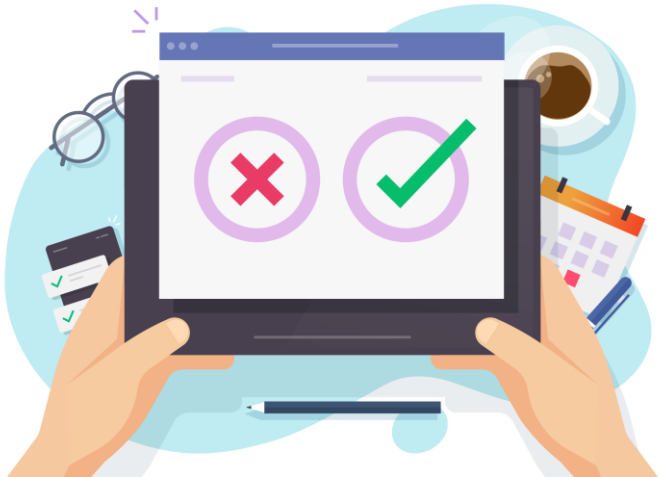
By the End of this Session, You will:

- Gain a solid understanding of the importance of data visualization in data science.
- Familiarize yourself with various types of visualizations and their applications.
- Learn about key libraries for data visualization in Python, including their features and use cases.
- Acquire knowledge about the components of a plot and how to create effective visualizations using Pandas and other libraries.

Poll Time

Q. Why is data visualization important in data science?

a) To enhance the aesthetics of data
b) To summarize and manipulate data efficiently
c) To communicate complex information effectively
d) To perform statistical analysis of data



Poll Time

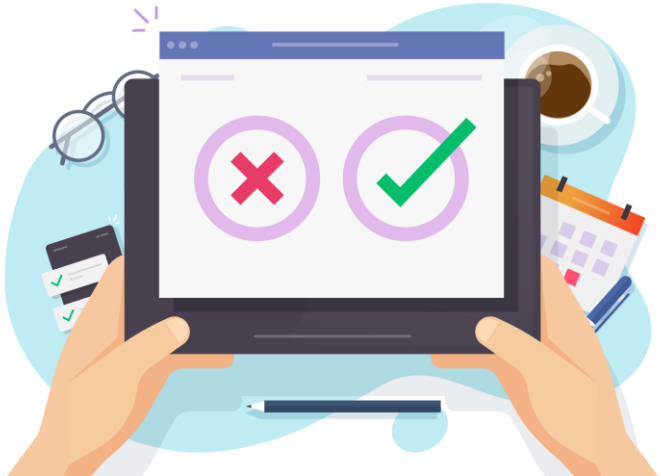
Q. Why is data visualization important in data science?

a) To enhance the aesthetics of data

b) To summarize and manipulate data efficiently

c) To communicate complex information effectively

d) To perform statistical analysis of data



Introduction to Data Visualization

Need for Visualization

1. **Data comprehension:** Visualization helps in understanding complex datasets by presenting information in a visual format. It enables us to identify patterns, trends, and outliers more easily than analyzing raw data.

2. **Communication and storytelling**: Visualizations are powerful tools for communicating insights and telling a compelling story with data. They help in conveying information effectively to both technical and non-technical audiences.

3. **Decision-making:** Visualizations aid in making informed decisions by providing a visual representation of data. They enable stakeholders to grasp the significance of different options and potential outcomes quickly.

Need for Visualization

4. **Exploring relationships**: Visualizations allow us to explore relationships and correlations between variables. By visually representing the data, we can identify connections and dependencies that may not be apparent in tabular form.

5. **Identifying insights and anomalies**: Visualizations help in spotting anomalies and outliers in data, facilitating the detection of errors or unusual patterns. They also assist in uncovering hidden insights and generating new hypotheses for further analysis.

Types of Visualizations

1. **Bar charts**: Used to compare and display categorical data using rectangular bars. They are effective for visualizing data across different categories or groups.

2. **Line charts**: Show the relationship between two numeric variables using continuous lines. They are useful for tracking trends and changes over time.

3. **Scatter plots**: Represent the relationship between two variables as individual data points. They help in identifying patterns, correlations, and outliers in the data.

Types of Visualizations

4. **Pie charts**: Display the proportion or percentage of each category in a dataset as a circular chart divided into slices. They are ideal for visualizing relative proportions or compositions.

5. **Heatmaps**: Present data in a grid format using color gradients. They are commonly used to show the magnitude, intensity, or relationships between two categorical variables.

6. **Histograms**: Visualize the distribution of a single numeric variable by dividing it into bins and showing the frequency of data points in each bin. Histograms are useful for understanding the shape and spread of data.

Key Libraries to be Used

1. **Matplotlib:** A versatile and widely-used plotting library that provides a wide range of plotting functions, including line plots, scatter plots, bar charts, histograms, and more. It serves as the foundation for many other visualization libraries.

2. **Seaborn:** Built on top of Matplotlib, Seaborn offers a higher-level interface for creating attractive statistical visualizations. It provides easy-to-use functions for creating complex visualizations like heatmaps, violin plots, and categorical plots.

3. **Pandas:** While primarily known for its data manipulation capabilities, Pandas also offers basic plotting functionality. It allows you to create simple visualizations directly from Pandas data structures, such as line plots, bar charts, histograms, and box plots.

Components of a Plot

1. **Axes**: The axes represent the coordinate system of the plot. They consist of two perpendicular lines, the x-axis and the y-axis, which provide a reference for positioning and scaling the data.
2. **Data Points:** Data points are the individual values or observations being plotted. They are represented as markers, dots, or other visual symbols on the plot.
3. **Lines or Curves**: Lines or curves connect data points to show the relationship or trend between them. They can represent a variety of relationships, such as a line of best fit, a trendline, or a curve describing a mathematical function.
4. **Labels and Titles**: Labels and titles provide descriptive information about the plot, including the axes labels, a title for the plot, and any additional annotations or explanations necessary for understanding the visualization.

Components of a Plot

5. **Legends:** Legends are used to explain the meaning of different elements in the plot, such as different categories or groups represented by different colors or markers.
6. **Gridlines**: Gridlines are horizontal and vertical lines that help in visually aligning and interpreting the data points. They assist in reading values from the plot accurately and provide a reference for scale and positioning.
7. **Annotations**: Annotations are additional text or graphical elements that provide supplemental information about specific data points or regions of interest in the plot. They help in highlighting important features or insights.

Demo – Basic Plots

Pop Quiz

Q. What is the difference between a line chart and a scatter plot?

a. A line chart uses connected points to represent the data, while a scatter plot uses individual points

b. A line chart can be used to show trends in the data, while a scatter plot cannot

c. A line chart is better for showing the relationship between two or more variables, while a scatter plot is better for showing how data changes over time

d. A line chart is better for showing the frequency of data, while a scatter plot cannot

Pop Quiz

Q. What is the difference between a line chart and a scatter plot?

a. A line chart uses connected points to represent the data, while a scatter plot uses individual points

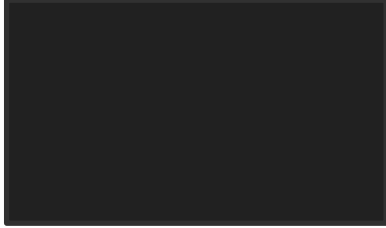
b. A line chart can be used to show trends in the data, while a scatter plot cannot

c. A line chart is better for showing the relationship between two or more variables, while a scatter plot is better for showing how data changes over time

a. A line chart is better for showing the frequency of data, while a scatter plot cannot

Subplots

Introduction to Subplots



1. **Multiple Plots in a Single Figure**: Subplots allow you to create multiple plots within a single figure window. Each subplot represents a separate visualization or chart, allowing you to display and compare different aspects of your data simultaneously.

2. **Grid Structure:** Subplots are organized in a grid structure, typically with rows and columns. This grid defines the layout of the subplots on the figure canvas. You can specify the size and position of each subplot within the grid.

3. **Enhanced Data Comparison:** Subplots enable you to compare different visualizations or variations of the same data side by side. This makes it easier to identify patterns, trends, and differences across different subsets of your data or different variables.

Introduction to Subplots

4. **Flexibility and Customization:** Subplots offer flexibility in terms of the number of rows and columns, allowing you to create various grid configurations. You can customize the size, aspect ratio, and spacing between subplots to achieve the desired layout and visual appeal.

5. **Integrated Data Exploration:** Subplots facilitate interactive data exploration by allowing you to interact with individual plots within the same figure. You can zoom in, pan, or apply specific operations to a particular subplot while keeping the context of the other subplots intact.

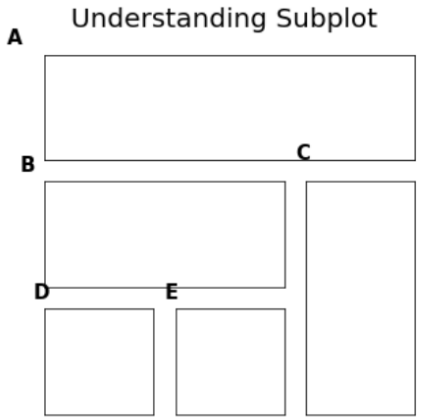
Customizing Subplots

- Use subplots() to create a grid of subplots.
- Use ax to access and customize individual subplots.
- Use gridspec() to create a complex subplot layout.
- Use plt.subplots_adjust() to adjust the spacing between subplots.

CODE

```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(2, 2, figsize=(5, 5))

ax1 = plt.subplot2grid((3, 3), (0, 0), colspan=3)
ax2 = plt.subplot2grid((3, 3), (1, 0), colspan=2)
ax3 = plt.subplot2grid((3, 3), (1, 2), rowspan=2)
ax4 = plt.subplot2grid((3, 3), (2, 0))
ax5 = plt.subplot2grid((3, 3), (2, 1))

axes = fig.get_axes()
texts = ['A', 'B', 'C', 'D', 'E']
for a, l in zip(axes, texts):
    a.annotate(l, xy=(-0.1, 1.1), xycoords="axes fraction", fontsize=15, weight='bold')
    a.set_xticks([])  # Remove x-axis ticks
    a.set_yticks([])  # Remove y-axis ticks

plt.suptitle('Understanding Subplot', fontsize=20)
plt.show()
```

OUTPUT



Pop Quiz

Q. What is the advantage of using subplots in EDA?

a. Subplots allow you to visualize multiple data sets in a single figure

b. Subplots allow you to compare multiple data sets over time

c. Subplots allow you to show how a single data set changes over time

d. All of the listed

Pop Quiz

Q. What is the advantage of using subplots in EDA?

a. Subplots allow you to visualize multiple data sets in a single figure

b. Subplots allow you to compare multiple data sets over time

c. Subplots allow you to show how a single data set changes over time

d. All of the listed

Demo - Creating and Customizing Subplots

Pop Quiz

Q. What is the difference between a subplot and a grid?

a. A subplot is a single graph within a figure, while a grid is a collection of subplots

b. A subplot is used to show how data changes over time, while a grid is used to show the relationship between two or more variables

c. A subplot is used to show the distribution of data, while a grid is used to show the frequency of data

d. There is no difference between a subplot and a grid

Pop Quiz

Q. What is the difference between a subplot and a grid?

a. A subplot is a single graph within a figure, while a grid is a collection of subplots

b. A subplot is used to show how data changes over time, while a grid is used to show the relationship between two or more variables

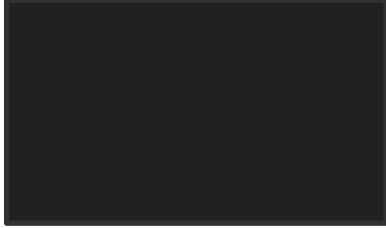
c. A subplot is used to show the distribution of data, while a grid is used to show the frequency of data

d. There is no difference between a subplot and a grid


Q&A

Introduction to Seaborn

Introduction to Seaborn



1. **High-level Data Visualization Library**: Seaborn is a Python data visualization library that provides a high-level interface for creating attractive and informative statistical graphics. It is built on top of Matplotlib and offers enhanced aesthetics and advanced statistical plotting capabilities.

2. **Beautiful Visualizations with Minimal Code**: Seaborn simplifies the process of creating visually appealing plots by providing ready-to-use themes and color palettes. With just a few lines of code, you can create stunning visualizations with pleasing aesthetics.

Introduction to Seaborn

3. **Statistical Plotting Made Easy:** Seaborn specializes in statistical plotting, offering a wide range of specialized plot types. It provides easy-to-use functions for creating visualizations such as box plots, violin plots, scatter plots, and regression plots. These plots enable effective exploration and communication of relationships, distributions, and comparisons within your data.

4. **Seamless Integration with Pandas**: Seaborn seamlessly integrates with the Pandas library, a popular data manipulation and analysis tool in Python. It allows you to easily create visualizations directly from Pandas DataFrames, streamlining the workflow for data exploration and analysis.

Pop Quiz

Q. Which of the following is a strength of Matplotlib?

a. It has a wide range of features and customization options

b. It is well-documented and has a large community of users

c. It is easy to learn and use

d. All of the above

Pop Quiz

Q. Which of the following is a strength of Matplotlib?

a. It has a wide range of features and customization options

b. It is well-documented and has a large community of users

c. It is easy to learn and use

d. All of the above

Demo - Creating a Basic Seaborn Plot

Pop Quiz

Q. Which of the following is a strength of Seaborn?

a. It provides a number of pre-defined styles for plots

b. It is easy to learn and use

c. It can be used to create interactive plots

d. All of the above

Pop Quiz

Q. Which of the following is a strength of Seaborn?

a. It provides a number of pre-defined styles for plots

b. It is easy to learn and use

c. It can be used to create interactive plots

d. All of the above

Demo - Bar Chart and Pie Chart

Pop Quiz

Q. Which type of chart is better for showing trends in data?

a. Bar chart

b. Line chart

c. Either bar chart or pie chart

d. It depends on the specific data sets

Pop Quiz

Q. Which type of chart is better for showing trends in data?

a. Bar chart

b. Line chart

c. Either bar chart or pie chart

d. It depends on the specific data sets

Demo – Histogram and Scatter Plots

Pop Quiz

Q. Which of the following is not a characteristic of a histogram?

a. It is a type of graph that shows the distribution of data

b. It uses a series of horizontal bars to represent the data

c. The bars are usually of equal width

d. The bars can be stacked on top of each other to show the relative frequency of data

Pop Quiz

Q. Which of the following is not a characteristic of a histogram?

a. It is a type of graph that shows the distribution of data

b. It uses a series of horizontal bars to represent the data

c. The bars are usually of equal width

d. The bars can be stacked on top of each other to show the relative frequency of data

Pop Quiz

Q. Which of the following is the correct syntax to plot a scatterplot using pandas?

a. pandas.plot(data, x='x', y='y')

b. pandas.scatterplot(data, x='x', y='y')

c. data.plot(x='x', y='y')

d. data.scatterplot(x='x', y='y')

Pop Quiz

Q. Which of the following is the correct syntax to plot a scatterplot using pandas?

a. pandas.plot(data, x='x', y='y')

b. pandas.scatterplot(data, x='x', y='y')

c. data.plot(x='x', y='y')

d. data.scatterplot(x='x', y='y')


Q&A

Activity 1

Pre-requisites:
Pandas Scatterplot and matplotlib.

Scenario:
A company wants to track product sales over time by month and see if there's a relationship between time and sales.

Data:
df = pd.DataFrame({ "month": ["January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December"], "number_of_products_sold": [100, 120, 150, 180, 200, 220, 240, 260, 280, 300, 320] })

Expected Outcome:
The final result should be a scatterplot with a title, labels, and a trend line if there is one.

Steps:
1. Import the pandas and matplotlib.pyplot libraries. Create a DataFrame with the data.
2. Plot a scatterplot of the x and y values using the DataFrame.plot() method.
3. Add a title to the plot.
4. Add labels to the x-axis and y-axis.
5. Add a trend line to the plot, if there is one.
6. Show the plot.

Activity 2

Pre-requisites:
Pandas Scatterplot and matplotlib.

Scenario:
A company wants to track the number of website visitors over time and see if there is any relationship between the number of visitors and the time of day.

Data:
df = pd.DataFrame({ "hour": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], "number_of_website_visitors": [100, 120, 150, 180, 200, 220, 240, 260, 280, 300, 320] })

Expected Outcome:
The final result should be a scatterplot with a title, labels, and a trend line if there is one.

Steps:
1. Import the pandas and matplotlib.pyplot libraries. Create a DataFrame with the data.
2. Plot a scatterplot of the x and y values using the DataFrame.plot() method.
3. Add a title to the plot.
4. Add labels to the x-axis and y-axis.
5. Add a trend line to the plot, if there is one.
6. Show the plot.

Summary

• Matplotlib and Seaborn are Python libraries for creating static, animated, and interactive visualizations.

• They provide a wide range of plotting functions for creating different types of plots, including line plots, bar charts, histograms, and scatter plots.

• They allow you to customize the appearance of your plots, including the colors, fonts, and styles of the axes, labels, and legends.

• They can be used to create interactive plots that allow users to zoom in and out, pan around, and hover over points to see more information.

Next Session:
Data Visualization – Case Study

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad

knowledgehut
upGrad

Data Visualization - Case Study

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the End of this Session, You Will:

- Learn the significance of data visualization in uncovering patterns and trends in data.
- Convert a CSV file into a table format and identify and remove null and duplicate values from the dataset.
- Visualize the correlation between different features using various plots.
- Draw a count plot of all the distinct classes of a feature and identify the relationship between two features.

Recap

Poll Time

Q. What is the significance of data visualization?

a. Data visualization can help identify patterns and trends in data

b. Data visualization can help understand the relationships between different variables

c. Data visualization can help communicate the findings to others

d. All of the listed

Poll Time

Q. What is the significance of data visualization?

a. Data visualization can help identify patterns and trends in data

b. Data visualization can help understand the relationships between different variables

c. Data visualization can help communicate the findings to others

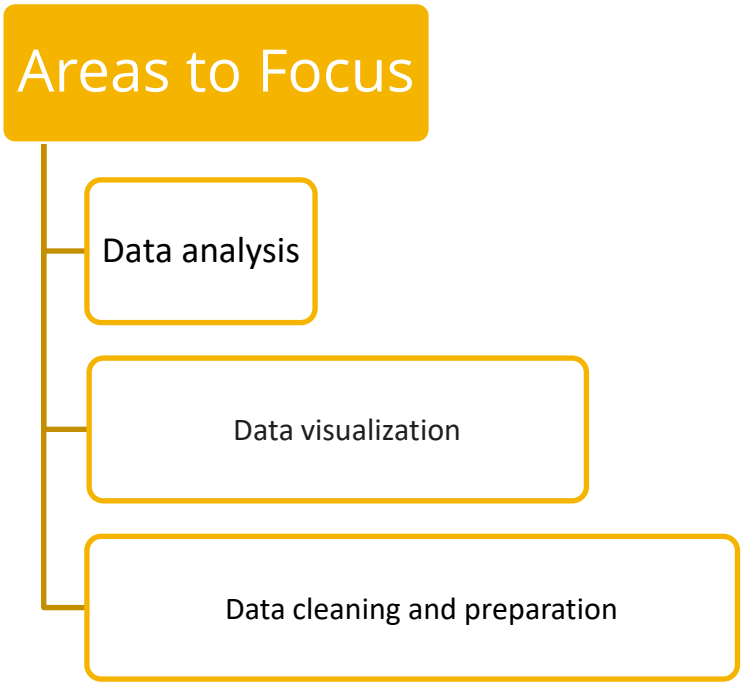
d. All of the listed

Case Study – Problem Statement

Problem Statement

- John owns a series of electronics stores and has gathered information on the sales of various smartphones.
- He is particularly interested in determining which phone has sold the most units among all the smartphones in his inventory.
- To achieve this goal, he requires a thorough analysis of the smartphone sales data he has collected.
- The dataset comprises several columns, including battery power, clock speed, internal memory, RAM size, touch screen, and others.
- By exploring this dataset, John hopes to uncover patterns and trends that could aid him in making informed business decisions.
- He has sought your assistance in visualizing this data, allowing him to gain greater insight into the sales performance of his smartphone inventory.

Areas to Focus



Poll Time

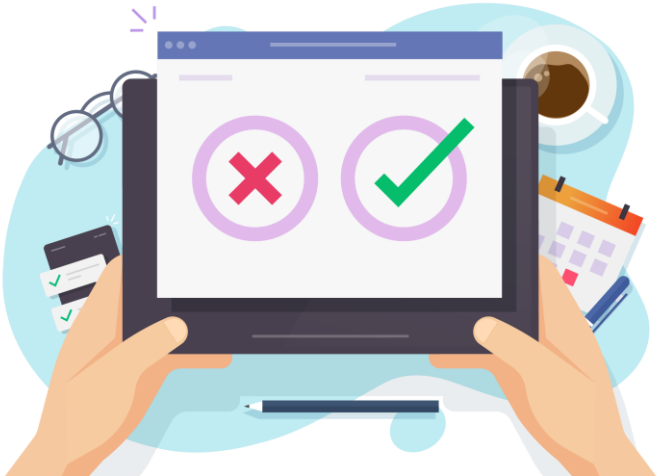
Q. What are some specific data visualization techniques that you could use to explore the data?

a. Bar charts

b. Line charts

c. Scatter plots

d. All of the listed



Poll Time

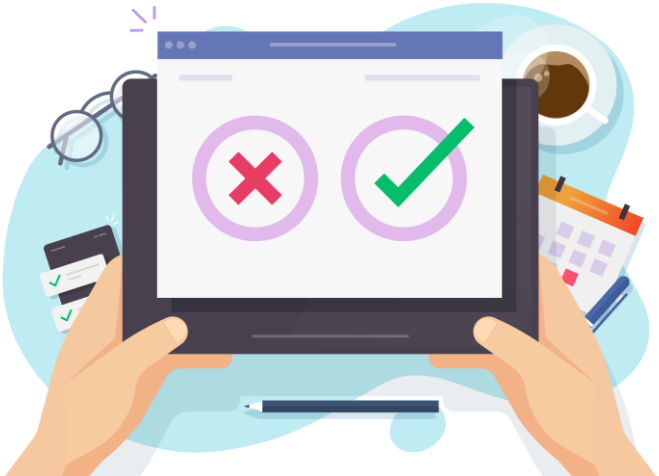
Q. What are some specific data visualization techniques that you could use to explore the data?

a. Bar charts

b. Line charts

c. Scatter plots

d. All of the listed



Understanding the Data

Sneak Peak into the Data

Physical Features

Feature	Description
battery_power	The total energy a battery can store at one time, measured in mAh.
m_dep	The depth of the smartphone, measured in centimeters.
mobile_wt	The weight of the smartphone, measured in grams.

Sneak Peak into the Data

Technical Features

Feature	Description
clock_speed	The speed at which the microprocessor executes instructions.
int_memory	The amount of internal memory in the smartphone, measured in gigabytes.
n_cores	The number of cores in the smartphone's processor.
Ram	The amount of random-access memory in the smartphone, measured in megabytes.
px_height	The height of the smartphone's screen, measured in pixels.
px_width	The width of the smartphone's screen, measured in pixels.

Sneak Peak into the Data

Communication Features

Feature	Description
Blue	Indicates whether or not the smartphone has Bluetooth.
four_g	Indicates whether or not the smartphone supports 4G.
three_g	Indicates whether or not the smartphone supports 3G.
touch_screen	Indicates whether or not the smartphone has a touch screen.
wifi	Indicates whether or not the smartphone has Wi-Fi.
price_range	Indicates the price range of the smartphone.

What are You Going to Build?

- Identify unnecessary columns from the dataset and remove them.
- Analyzing the count plot of all the distinct classes of the price_range feature.
- Determining the relationship between the price of the devices and their battery capacities.
- Identifying unnecessary columns from the dataset and removing them.
- Visualizing the correlation between all the features.
- Analyzing various distributions like the distribution of smartphones by price range.

Poll Time

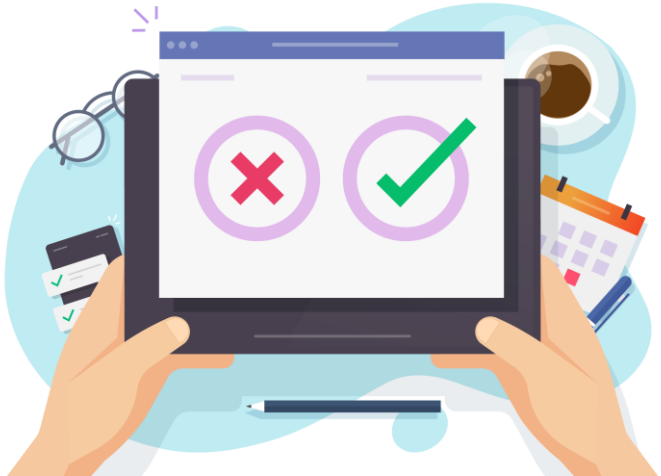
Q. Which of the following is a way to visualize the correlation between all the features?

a. Bar chart

b. Line chart

c. Scatter plot

d. Heatmap



Poll Time

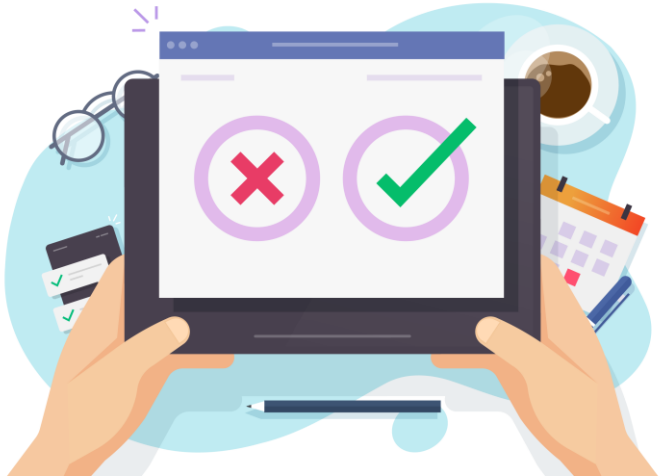
Q. Which of the following is a way to visualize the correlation between all the features?

a. Bar chart

b. Line chart

c. Scatter plot

d. Heatmap




Q&A

Hands-on: Case Study Questions

Poll Time

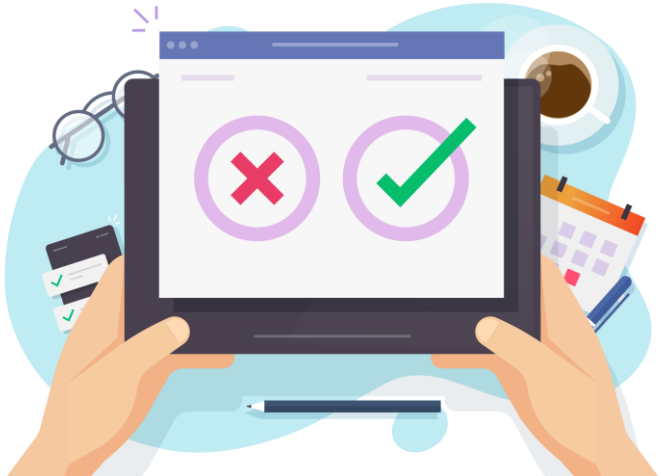
Q. Which of the following is a way to identify outliers in a dataset?

a. Bar chart

b. Line chart

c. Scatter plot

d. Histogram



Poll Time

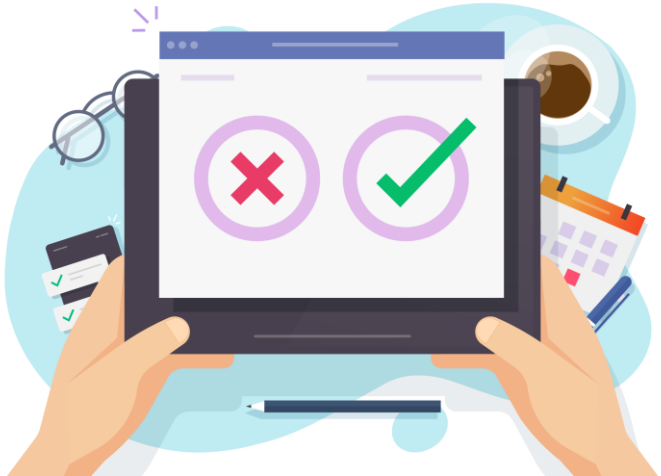
Q. Which of the following is a way to identify outliers in a dataset?

a. Bar chart

b. Line chart

c. Scatter plot

d. Histogram




Q&A

Summary

- Data visualization can help us identify patterns and trends in data.
- Data cleaning is an important step in any data analysis process.
- Data analysis can be used to answer a variety of questions about data.
- Data cleaning can help to ensure that data is accurate and reliable.

Activity 1

Pre-requisites:
Basics of data visualization, such as bar charts, line charts, and scatter plots.

Scenario:
You are a marketing manager for a smartphone company. You want to create a visualization that shows the sales of your company's smartphones over the past year.

Expected outcome:
Your visualization should show the sales of your company's smartphones over the past year. You should use a variety of data visualization techniques to highlight different trends in sales.

Steps:
1.Create a Data Frame as per the given data.
2.Choose the data visualization techniques that you want to use.
3.Create the visualization and note the inference.

Smartphone Model	Sales (Units)	Month
iPhone 13 Pro Max	100,000	January
iPhone 13 Pro	80,000	January
iPhone 13	60,000	January
Samsung Galaxy S22 Ultra	50,000	January
Samsung Galaxy S22+	40,000	January
Samsung Galaxy S22	30,000	January

Activity 2

Pre-requisites:
Basics of data visualization, such as bar charts, line charts, and scatter plots.

Scenario:
You are a product manager for a smartphone company. You want to create a visualization that shows the customer satisfaction with your company's smartphones.

Expected outcome:
Your visualization should show the customer satisfaction with your company's smartphones. You should use a variety of data visualization techniques to highlight different areas of customer satisfaction.

Steps:
1.Create a Data Frame as per the given data.
2.Choose the data visualization techniques that you want to use.
3.Create the visualization and note the inference.

Customer Satisfaction Rating	Number of Customers
5	100
4	200
3	300
2	400
1	500

Session Feedback

Next Session:
Web Scraping and Exploratory Data Analysis

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad