Data Analysis

Data Visualization

Zeham Management Technologies BootCamp by SDAIA

July 24th, 2024

Introduction to Data Visualization

Let’s start together…

Agenda

- Data Visualization
- Benefits of Visualization
- Types of Data Analysis
- Charts Definitions
- Charts Usage
- Python Visualization Libraries
- Python interactive Visualizations
- BI Solutions

Data Visualization

Definition:

- 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.

Use:

- The primary goal of data visualization is to communicate information clearly and efficiently to users, enabling them to understand complex data easily. It helps in making data-driven decisions by revealing insights that might not be apparent from raw data.

Data Visualization Benefits

Easier To understand

Identifies Relationships

Faster Decision Making

Better for monitoring

Better for monitoring

Detect Errors And Measures risks

Types of Data Analysis

Univariate Analysis

Definition:

- Analysis of a single variable.
- Purpose: To describe and summarize data..

Key Metrics:

- Mean, median, mode
- Variance, standard deviation
- Distribution shape (e.g., skewness, kurtosis)

Visualization Techniques:

- Histograms
- Box plots
- Bar charts

Bivariate Analysis

Definition:

- Analysis of the relationship between two variables.
- Purpose: To find correlations and understand the connection between variables.

Key Metrics:

- Correlation coefficient
- Covariance

Visualization Techniques:

- Scatter plots
- Line Graphs Heatmaps

Multivariate Analysis

Definition:

- Analysis involving more than two variables.
- Purpose: To understand complex interactions and relationships.

Key Techniques:

- Multiple regression
- Principal Component Analysis (PCA)
- Cluster analysis

Visualization Techniques:

- Pair plots
- 3D scatter plots
- Parallel coordinates plot
- Heatmaps with multiple variables

Line Chart

Charts

Definition:

- Connects individual data points with line segments.

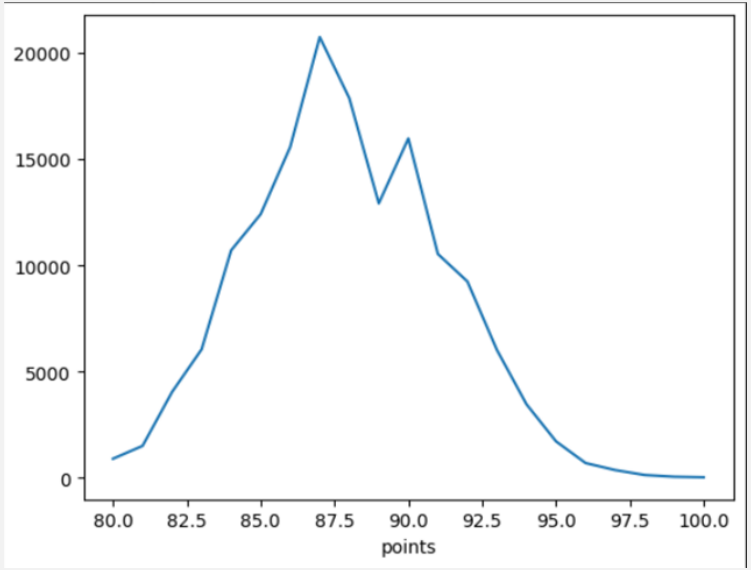
Use:

- Ideal for showing trends and changes over time for one or more groups.

Data Types:

- Best with continuous data, particularly effective for time series analysis.

```
1  reviews['points'].value_counts().sort_index().plot.line()
```



Pie Chart

Charts

Definition:

- Represents data as slices of a pie, with slice sizes proportional to the part-to-whole relationships.

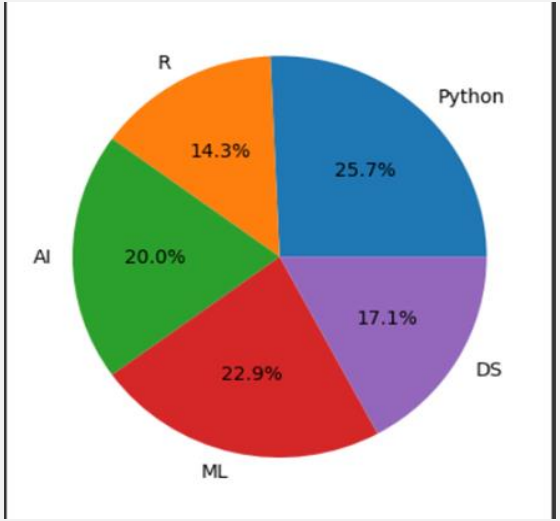
Use:

- Useful for displaying the composition of a dataset in a few categories.

Data Types:

- Best with Percentages or Categorical data counts

```
plt.pie(class1_student,labels=classes,autopct="%0.1f%%")
plt.show()
```



Histogram Chart

Charts

Definition:

- A bar chart representing the frequency distribution of a single continuous variable.

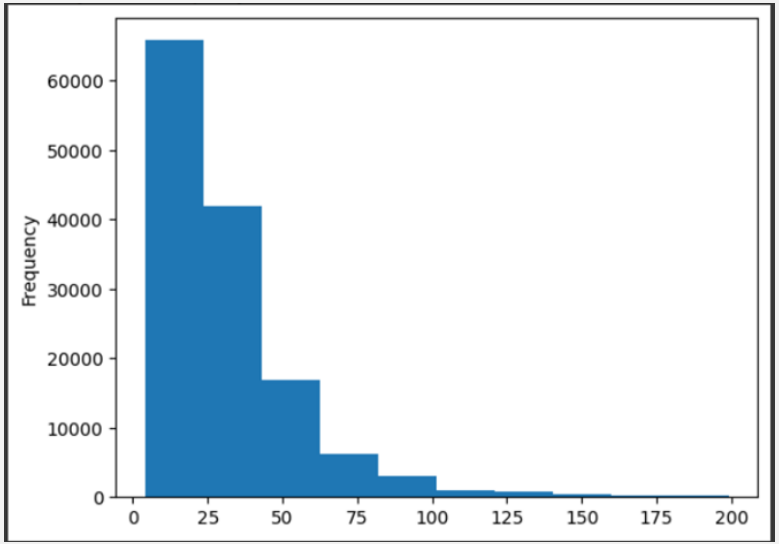
Use:

- Excellent for visualizing the distribution, skewness, and modality of the data.

Data Types:

- Continuous data, particularly for analyzing the shape of distributions..

```
1  reviews[reviews['price'] < 200]['price'].plot.hist()
```



Bar Chart

Charts

Definition:

- Uses horizontal or vertical bars to show comparisons among categories.

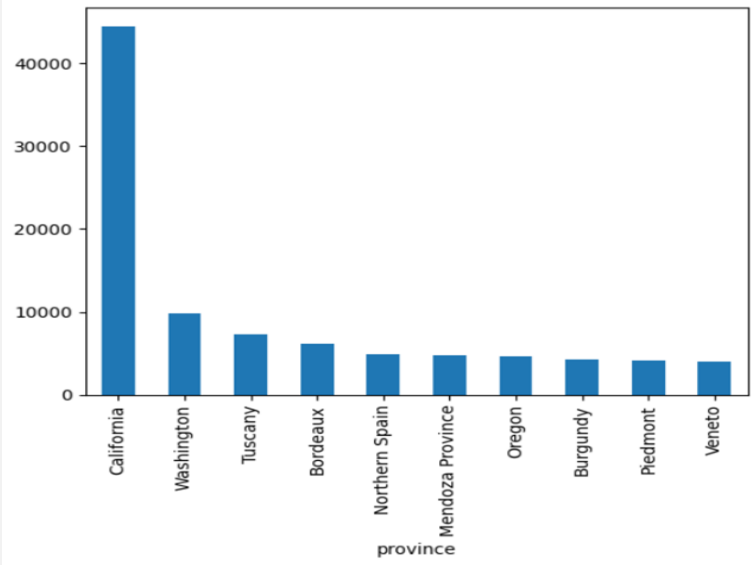
Use:

- Effective for comparing quantities across different categories.

Data Types:

- Categorical data on the x-axis and continuous data on the y-axis often used to count occurrence of categories or to show prices group by country for example.

```
(reviews['province'].value_counts().head(10) /
len(reviews)).plot.bar()
```



Univariate Analysis Example

Scatter Plot

Charts

Definition:

- Displays values for typically two variables for a set of data using Cartesian coordinates.

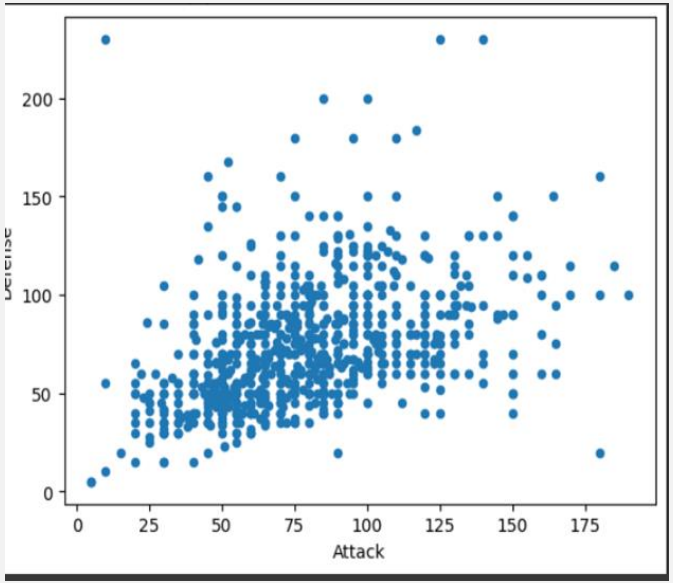
Use:

- Identifies the relationship, trend, or correlation between two variables.

Data Types:

- Continuous data; suitable for regression and correlation analysis.

```
1  pokemon.plot.scatter(x='Attack', y='Defense')
```



Bivariate Analysis Example

Heatmap

Charts

Definition:

- Uses color coding to represent complex data matrices and highlight variances.

Use:

- Excellent for detecting relationships, correlations, or areas of intensity.

Data Types:

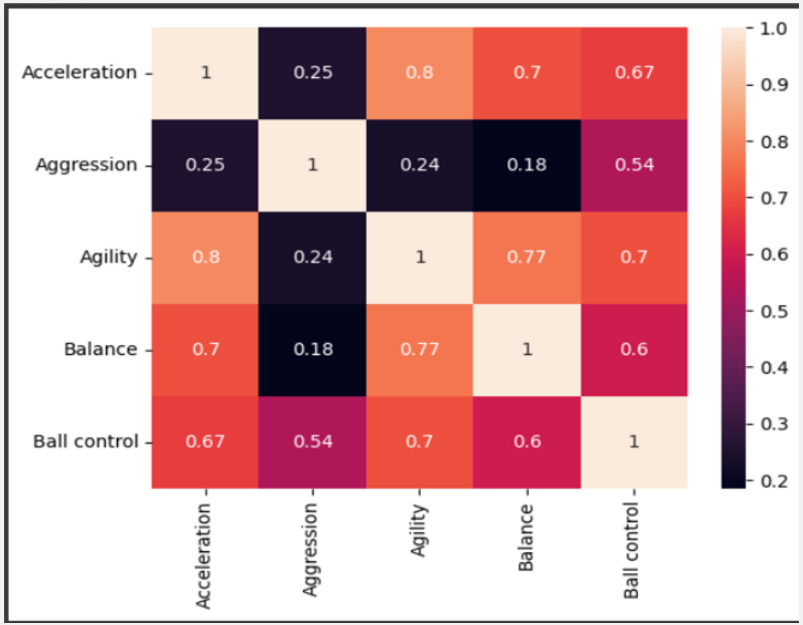
- Continuous or categorical data, useful for cross-tabulations or correlations.

Heatmap

Charts

```
f = (
    footballers.loc[:, ['Acceleration', 'Aggression', 'Agility', 'Balance', 'Ball control']]
        .applymap(lambda v: int(v) if str.isdecimal(v) else np.nan)
        .dropna()
).corr()

sns.heatmap(f, annot=True)
```



HexBin plot

```
pokemon.plot.hexbin(x='Attack', y='Defense', gridsize=15)
```

Definition:

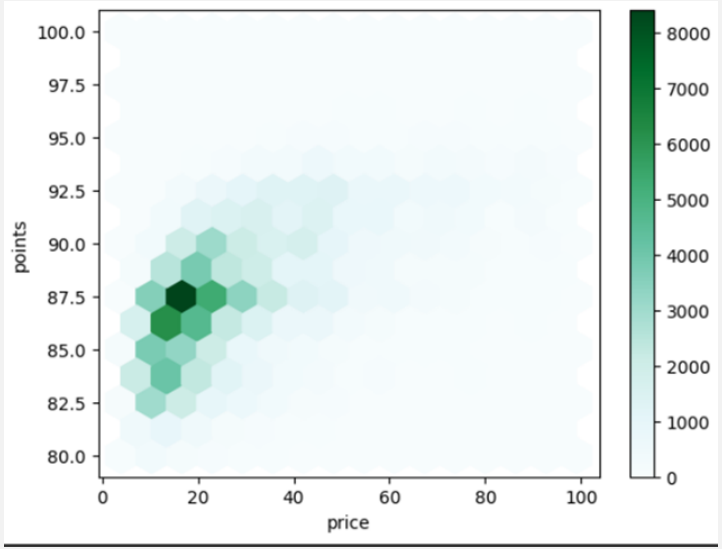
Combines scatter and density plots using hexagonal bins to show data distribution.

Use:

Visualizes data density:
Highlights clusters and trends.
Mitigates overplotting issues.

Data Types:

Continuous data:
Ideal for large datasets with two variables.
Useful for comparing distributions across groups.



Bivariate Analysis Example

Radar Chart

Charts

Definition:

- Displays multivariate data in the form of a two-dimensional chart.

Use:

- Compares three or more quantitative variables represented on axes starting from the same point.

Data Types:

- Continuous data, ideal for displaying performance metrics across multiple categories..

Violin Plot

Charts

Definition:

- Combines elements of box plots and density plots, showing data distribution and probability density.

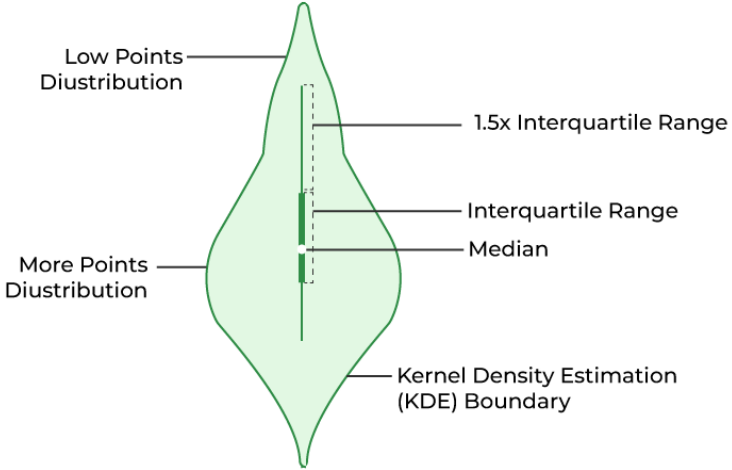
Use:

- Offers a deeper view of data distribution, showing peaks and valleys within the data.

Data Types:

- Continuous data, ideal for comparing distributions across categories or groups.

```
1  sns.violinplot(data=distributions, orient='h',
   bw=0.05, cut=0);
2
```



Let’s Practice

Notebook Path:

4.Data Visualization techniques with python/LAB/Data_Visualization_Tutorial.ipynb

Python Visualization Libraries

Matplotlib

Python Visualization Libraries

Matplotlib is a widely used Python library for creating a variety of visualizations, including static, animated, and interactive plots.

Key Features include:

- **Versatility:** Supports numerous plot types like line charts, bar charts, and histograms.
- **Customization:** Offers detailed customization for almost all aspects of a plot.
- **User-Friendly:** Simple to use for beginners while providing advanced features for complex visualizations.
- **Integration:** Integrates seamlessly with NumPy and pandas, making it a staple in Python data analysis.
- **Community Support:** Benefits from extensive documentation and a large user community.

Seaborn

Python Visualization Libraries

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Key Features include:

- **Ease of Use:** Simplifies creating complex visualizations.
- **Pandas Integration:** Optimized for pandas data structures.
- **Improved Aesthetics:** Offers attractive themes and color palettes.
- **Statistical Tools:** Built-in functions for adding statistical details to plots.
- **Variety of Plots:** Supports diverse plots like scatter, violin, and heatmaps.
- **Customizable:** Allows detailed control while being easier than Matplotlib.

Plotly

Python interactive Visualizations Libraries

Plotly is an interactive, open-source graphing library for Python that enables users to create visually appealing, interactive, and web-friendly charts and dashboards.

It supports a wide array of chart types, including more advanced visualization types like 3D plots and geographical maps.

Key Features include:

- **Interactivity:** Supports zoom, pan, and hover effects.
- **Wide Range of Plots:** Offers diverse plot types including 3D plots and maps.
- **Web Integration:** Easy embedding in web apps.
- **Compatibility:** Works well with Pandas and NumPy.
- **User-friendly:** Simplifies creating complex visualizations.
- **Customization:** Extensive options for plot customization.

Bokeh

Python interactive Visualizations Libraries

Bokeh is designed for creating interactive visualizations for modern web browsers. It offers a powerful and flexible toolkit for producing dynamic plots that can interact with large datasets or streaming data.

Key Features include:

- **Interactivity:** Features zoom, pan, and selection tools.
- **Versatility:** Offers a wide range of plot types.
- **Streaming Data:** Handles real-time data for dynamic updates.
- **Integration:** Works well with data science and web frameworks.
- **Customization:** Extensive appearance and functionality options.
- **Server Capability**: Includes a server for interactive web apps.

BI Solutions

Business Intelligence (BI) solutions are software tools that help companies analyze data to improve decision-making.

They include a range of applications and methods for collecting, preparing, analyzing data, and generating reports, dashboards, and visualizations.

Dashboards in Business Intelligence (BI):

- **Real-Time Visuals:** Charts and graphs for instant data and KPI analysis.
- **Accessibility:** Simplifies complex data for clear performance insights.
- **Customization:** Adapts dashboards for specific metrics and user requirements.
- **Data-Driven Decisions:** Emphasizes key trends for strategic planning and improvements.
- **Efficiency:** Quick, essential data access to streamline decision-making.

Thank you

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority