DATA SCIENCE

Code Kids Python L5

Term Definition & Glossary

pandas - A Python library for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series. |

DataFrame - A two-dimensional, size-mutable, potentially heterogeneous tabular data structure with labelled axes (rows and columns). It's a primary data structure used in pandas. |

Series - A one-dimensional labeled array capable of holding any data type. A Series is essentially a single column of a DataFrame in pandas. |

matplotlib - A Python library used for creating static, interactive, and animated visualizations. It is commonly used in data science for making graphs and charts. |

groupby() - A method used to group large amounts of data and compute operations on these groups. It can be used to summarize data statistically. |

loc[] - A function used for label-based indexing to access a group of rows and columns by labels or a boolean array. |

iloc[] - Similar to `loc[]`, but instead uses integer-based indexing to access data, allowing you to retrieve rows and columns by integer positions. |

head() - A method used to return the first 'n' rows of a DataFrame. Useful for getting a quick snapshot of the dataset. |

tail() - A method used to return the last 'n' rows of a DataFrame. This is useful for checking the last few entries in the data. |

isna() or **isnull()** - Methods used to detect missing values.

dropna() - A method used to remove missing values from a DataFrame or Series. It can remove all rows that contain a missing value. |
fillna() - A method used to fill NA/NaN values using a specified method, like 'ffill' for forward filling or a scalar value. |

csv - Stands for Comma-Separated Values. It's a format for saving tabular data in plain text, where each line corresponds to a data record, and each record consists of fields delimited by commas. |

read_csv() - A function in pandas used to read a CSV file into a DataFrame. It is highly customizable with respect to handling different delimiters, missing value representations, and column types. |

plot - Used to create a variety of plots, including line plots, bar charts, histograms, and more. In pandas, this is often used as a quick way to visualize data directly from DataFrames. |

bar chart - A type of graph that represents data with rectangular bars with lengths proportional to the values that they represent. Often used to compare various categories. |

pie chart - A circular statistical graphic, divided into slices to illustrate numerical proportion. Each slice's size is proportional to the quantity it represents. |