COURSE GLOSSARY

Exploratory Data Analysis in Python

.agg (aggregate): A method used after groupby or on a DataFrame to apply one or more aggregation functions (e.g., mean, sum, std) to columns and return summary results

.astype: A pandas Series or DataFrame method used to convert a column or columns to a different data type, such as int, float, or datetime

.dtypes: A pandas DataFrame attribute that lists the data type for each column, useful for checking whether columns are numeric, object, datetime, etc.

.info: A pandas DataFrame method that prints a concise summary of the dataset, including the number of non-null entries per column, data types, and memory usage

binwidth: A parameter used in histogram plotting that specifies the width of each bin, controlling the level of detail shown in the distribution

boxplot: A visualization that summarizes a numeric distribution by showing the median, quartiles, and potential outliers as points beyond whiskers

correlation (Pearson): A numeric measure ranging from -1 to 1 that quantifies the strength and direction of a linear relationship between two numeric variables

dropna: A pandas method that removes rows or columns containing missing values, with options to drop based on subsets or thresholds

Exploratory Data Analysis (EDA): The process of inspecting, cleaning, visualizing, and summarizing a dataset to discover patterns, detect anomalies, check assumptions, and generate hypotheses for further analysis

groupby: A pandas operation that groups rows by values in one or more columns to compute aggregate statistics within each group

heatmap: A colored matrix visualization that displays pairwise values such as correlation coefficients, with annotations and a color scale for quick pattern recognition

histogram: A visual representation of the distribution of a numeric variable that divides values into bins and plots the count or frequency in each bin

imputation: The practice of filling missing values with estimated values such as the mean, median, mode, or subgroup-specific summaries to retain observations for analysis

IQR (Interquartile Range): The difference between the 75th percentile (Q3) and the 25th percentile (Q1), used to measure spread and to define outlier thresholds

isna: A pandas method that returns a boolean mask indicating which values are missing (NA/NaN) in a Series or DataFrame

missing data: Data entries that are absent or recorded as NA/NaN, which can bias analyses and must be handled through removal, imputation, or other strategies

outlier: An observation that lies far outside the expected range of a variable, often defined as below Q1 - 1.5*IQR or above Q3 + 1.5*IQR, which can distort summary statistics

pd.read_csv: A pandas function that reads data from a CSV file and returns a DataFrame, with options to parse dates, set column types, and handle missing values

pd.to_datetime: A pandas function that converts strings or separate year/month/day columns into datetime objects, enabling time-based indexing and feature extraction

scatterplot: A two-dimensional plot that shows the relationship between two numeric variables by placing points at their (x, y) coordinates to reveal trends or clusters

Series: A one-dimensional labeled array in pandas representing a single column of data within a DataFrame or a standalone sequence of values

transform: A pandas method that computes group-wise statistics and returns an aligned Series or DataFrame of the same shape as the original, enabling per-row assignments of aggregated values

value_counts: A pandas Series method that returns the frequency counts of unique values in a categorical column, optionally normalized to proportions