# Terms and definitions from all courses

## #

**\_\_init\_\_()**: A reserved method in Python classes known as a constructor **\_\_str\_\_()**: A special string method for constructors to tell Python what to display when the print function is called on a particular instance of a class

**%%time**: A magic command that provides the runtime of the cell it’s entered in to

## 

## A

**A/B testing**: A way to compare two versions of something to find out which version performs better

**Absolute values**: (Refer to **observed values**)

**Accuracy**: Refers to the proportion of data points that were correctly categorized

**Action**: A Tableau tool to help an audience interact with a visualization or dashboard by allowing control of selection

**Active listening**: Refers to allowing team members, bosses, and other collaborative stakeholders to share their own points of view before offering responses

**AdaBoost**: (Refer to **adaptive boosting**)

**Adaptive boosting**: A boosting methodology where each consecutive base learner assigns greater weight to the observations incorrectly predicted by the preceding learner **Addition rule** (for **mutually exclusive events**): The concept that if the events A and B are mutually exclusive, then the probability of A or B happening is the sum of the probabilities of A and B

**Adjusted R²**: A variation of R**²** that accounts for having multiple independent variables present in a linear regression model

**Affinity**: The metric used to calculate the distance between points/clusters **Agglomerative clustering**: A clustering methodology that works by first assigning every point to its own cluster, then progressively combining clusters based on intercluster distance **Aggregate information**: Data from a significant number of users that has eliminated personal information

**Algorithm**: A set of instructions for solving a problem or accomplishing a task **Analysis of Variance (ANOVA):** A group of statistical techniques that test the difference of means between three or more groups

**Analytics Team Manager**: A data professional who supervises analytical strategy for an organization, often managing multiple groups **Analyze stage**: Stage of the PACE workflow where the necessary data is acquired from primary and secondary sources and then cleaned, reorganized, and analyzed **ANCOVA (Analysis of Covariance)**: A statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates, or variable(s) irrelevant to the test

**Append()**: A method that adds an element to a sequence **Array**: An ordered collection of items of a single data type **Array()**: A function for converting input to an array **Artificial intelligence (AI)**: Refers to computer systems able to perform tasks that normally require human intelligence

**Assignment**: The process of storing a value in a variable **Attributes**: Descriptive parameters of an object or class, usually in a list or dictionary

**Average**: The distance between each cluster’s centroid and other clusters’ centroids

## 

## B

**Backward elimination**: A stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

**Bagging**: A technique used by certain kinds of models that use ensembles of base learners to make predictions; refers to the combination of bootstrapping and aggregating

**Base learner**: Each individual model that comprises an ensemble **Bayes’ rule**: (Refer to **Bayes’ theorem**)

**Bayes’ theorem**: An equation that can be used to calculate the probability of an outcome or class, given the values of predictor variables

**Bayesian inference**: (Refer to **Bayesian statistics**)

**Bayesian statistics**: A powerful method for analyzing and interpreting data in modern data analytics; also referred to as Bayesian inference

**Best fit line**: The line that fits the data best by minimizing some loss function or error

**Bias**: In data structuring, refers to organizing data results in groupings, categories, or variables that are misrepresentative of the whole dataset

**Bias-variance trade-off**: Balance between two model qualities, bias and variance, to minimize overall error for unobserved data

**Bin**: A segment of data that groups values into categories

**Binning**: Grouping continuous values into a smaller number of categories, or intervals **Binomial distribution**: A discrete distribution that models the probability of events with only two possible outcomes: success or failure

**Binomial logistic regression**: A technique that models the probability of an observation falling into one of two categories, based on one or more independent variables

**Binomial logistic regression linearity assumption**: An assumption stating that there should be a linear relationship between each X variable and the logit of the probability that Y equals one

**Black-box model**: Any model whose predictions cannot be precisely explained **Boolean data**: A data type that has only two possible values, usually true or false **Boosting**: A technique that that builds an ensemble of weak learners sequentially, with each consecutive learner trying to correct the errors of the one that preceded it **Bootstrapping**: Refers to sampling with replacement

**Box plot**: A data visualization that depicts the locality, spread, and skew of groups of values within quartiles

**Branching**: The ability of a program to alter its execution sequence **Business Intelligence Analyst**: (Refer to **Business Intelligence Engineer**) **Business Intelligence Engineer**: A data professional who uses their knowledge of business trends and databases to organize information and make it accessible; also referred to as a Business Intelligence Analyst

## 

## C

**Categorical data**: Data that is divided into a limited number of qualitative groups **Categorical variables**: Variables that contain a finite number of groups or categories **Causation**: Describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Central Limit Theorem**: The idea that the sampling distribution of the mean approaches a normal distribution as the sample size increases

**Centroid**: The center of a cluster determined by the mathematical mean of all the points in that cluster

**Chi-squared (χ²) Goodness of Fit Test**: A hypothesis test that determines whether an observed categorical variable follows an expected distribution

**Chi-squared (χ²) Test for Independence**: A hypothesis test that determines whether or not two categorical variables are associated with each other

**Chief Data Officer**: An executive-level data professional who is responsible for the consistency, accuracy, relevancy, interpretability, and reliability of the data a team provides

**Child node**: A node that is pointed to from another node

**Class imbalance**: When a dataset has a predictor variable that contains more instances of one outcome than another

**Classes**: Code templates for creating objects

**Classical probability**: A type of probability based on formal reasoning about events with equally likely outcomes

**Cleaning**: The process of removing errors that might distort your data or make it less useful; one of the six practices of EDA

**Cluster random sample**: A probability sampling method that divides a population into clusters, randomly selects certain clusters, and includes all members from the chosen clusters in the sample

**Collaborative filtering**: A technique used by recommendation systems to make comparisons based on who else liked the content

**Collective outliers**: A group of abnormal points, following similar patterns and isolated from the rest of the population

**Comparator**: An operator that compares two values and produces Boolean values (true or false)

**Complement of an event**: In statistics, refers to an event not occuring **Complement rule**: A concept stating that the probability that event A does not occur is one minus the probability of A

**Complete**: The maximum pairwise distance between clusters **Composition**: Refers to defining attributes and methods at the instance level to have a more differentiated relationship between objects in the same class

**Concatenation**: Refers to building longer strings out of smaller strings **Conditional probability**: Refers to the probability of an event occurring given that another event has already occurred

**Conditional statement**: A section of code that directs the execution of programs **Confidence band**: The area surrounding a line that describes the uncertainty around the predicted outcome at every value of X

**Confidence interval**: A range of values that describes the uncertainty surrounding an estimate

**Confidence level**: A measure that expresses the uncertainty of the estimation process

**Confusion matrix**: A graphical representation of how accurate a classifier is at predicting the labels for a categorical variable

**Construct stage**: Stage of the PACE workflow where data models and machine learning

algorithms are built, interpreted, and revised to uncover relationships within the data and help unlock insights from those relationships

**Constructor**: A special method to add values to an instance in object creation

**Content-based filtering**: A technique used by recommendation systems to make comparisons based on attributes of content

**Contextual outliers**: Normal data points under certain conditions but become anomalies under most other conditions

**Continuous random variable**: A variable that takes all the possible values in some range of numbers

**Continuous**: A mathematical concept indicating that a measure or dimension has an infinite and uncountable number of outcomes

**Continuous variables**: Variables that can take on an infinite and uncountable set of values

**Convenience sample**: A non-probability sampling method that involves choosing members of a population that are easy to contact or reach

**Correlation**: Measures the way two variables tend to change together **Cross-validation**: A process that uses different portions of the data to test and train a model on different iterations

**CSV file**: A simple text file that can be easy to import or store in other softwares, platforms, and databases

**Customer churn**: The business term that describes how many and at what rate customers stop using a product or service, or stop doing business with a company

**Cut()**: A method that separates array elements into different bins

## 

## D

**Data anonymization**: The process of protecting people's private or sensitive data by eliminating PII

**Data cleaning**: The process of formatting data and removing unwanted material

**Data engineer**: A data professional who makes data accessible, ensures data ecosystems offer reliable results, and manages infrastructure for data across enterprises

**Data ethics**: Well-founded standards of right and wrong that dictate how data is collected, shared, and used

**Data governance**: A process for ensuring the formal management of a company’s data assets

**Data professional**: Any individual who works with data and/or has data skills

**Data science**: The discipline of making data useful

**Data scientist**: A data professional who works closely with analytics to provide meaningful insights that help improve current business operations

**Data source**: The location where data originates

**Data stewardship**: The practices of an organization that ensures that data is accessible, usable, and safe

**Data structure** : A collection of data values or objects that contains different data types **Data type**: An attribute that describes a piece of data based on its values, its programming language, or the operations it can perform

**Data visualization**: A graph, chart, diagram, or dashboard that is created as a representation of information

**Database (DB) file**: A file type used to store data, often in tables, indexes, or fields **Dataframe**: A two-dimensional data-structure organized into rows and columns **DBSCAN**: A clustering methodology that searches data space for continuous regions of high density; stands for “density-based spatial clustering of applications with noise”

**Debugging**: Troubleshooting, or searching for errors in a script or program

**Decision node**: A node of the tree where decisions are made **Decision tree**: A flowchart-like structure that uses branching paths to predict the outcomes of events, or the probability of certain outcomes

**Deduplication**: The elimination or removal of matching data values in a dataset **Def**: A keyword that defines a function at the start of the function block **Dependent events**: The concept that two events are dependent if one event changes the probability of the other event

**Dependent variable (Y)**: The variable a given model estimates **Describe()**: A function that returns the statistical summary of a dataframe or series, including mean, standard deviation, and minimum and maximum column values. **Descriptive statistics**: A type of statistics that summarizes the main features of a dataset **Dict()**: A function used to create a dictionary

**Dictionaries**: Data structures that consist of a collection of key-value pairs **Difference()**: A method that finds the elements present in one set but not the other **Dimensions**: Qualitative data values used to categorize and group data to reveal details about it

**Discovering**: The process data professionals use to familiarize themselves with the data so they can start conceptualizing how to use it; one of the six practices of EDA **Discrete features**: Features with a countable number of values between any two values **Discrete random variable**: A variable that has a countable number of possible values **Discrete**: A mathematical concept indicating that a measure or dimension has a finite and countable number of outcomes

**distance\_threshold**: A hyperparameter in agglomerative clustering models that determines the distance above which clusters will not be merged

**Docstring**: (Refer to **documentation string**)

**Documentation**: An in-depth guide that is written by the developers who created a package that features very specific information on various functions and features **Documentation string**: A group of text that explains what a method or function does; also referred to as a “docstring”

**Dot notation**: A type of notation that allows access to an object’s attributes and methods **Downsampling**: The process of removing some observations from the majority class, making it so they make up a smaller percentage of the dataset than before **Dummy variables**: Variables with values of 0 or 1 that indicate the presence or absence of something

**Dynamic value**: A value the user inputs or the output of a program, an operation, or a function

## 

## E

**Econometrics**: A branch of economics that uses statistics to analyze economic problems **Edge computing:** A way of distributing computational tasks over a bunch of nearby processors (i.e., computers) that is good for speed and resiliency and does not depend on a single source of computational power

**ELIF statement**: A statement that allows for more than two possible conditions in code **ELSE statement**: A statement that sets a piece of code to run only when the condition of the IF statement is false

**Empirical probability**: A type of probability based on experimental or historical data **Empirical rule**: A concept stating that the values on a normal curve are distributed in a regular pattern, based on their distance from the mean

**Ensemble learning**: Refers to building multiple models and aggregating their predictions **Ensembling**: (Refer to **ensemble learning**)

**Enumerate()**: A built-in function that iterates through a sequence and tracks each element and its place in the index

**eps** (Epsilon): In DBSCAN clustering models, a hyperparameter that determines the radius of a search area from any given point

**Errors**: In a regression model, the natural noise assumed to be in a model

**Execute stage**: Stage of the PACE workflow where a data professional will present findings with internal and external stakeholders, answer questions, consider different viewpoints, and make recommendations

**Explanatory variable**: (Refer to **independent variable**)

**Explicit conversion**: Refers to when users convert the data type of an object to a required data type

**Exploratory data analysis (EDA)**: The process of investigating, organizing, and analyzing datasets and summarizing their main characteristics, often by employing data wrangling and visualization methods; the six main practices of EDA are: discovering, structuring, cleaning, joining, validating, and presenting

**Expression**: A combination of numbers, symbols, or other variables that produce a result when evaluated

**Extra Sum of Squares F-test**: Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model **Extracting**: The process of retrieving data out of data sources for further data processing

**Extrapolation**: A model’s ability to predict new values that fall outside of the range of values in the training data

## F

**F1-Score**: The harmonic mean of precision and recall

**False positive**: A test result that indicates something is present when it really is not

**Feature engineering**: The process of using practical, statistical, and data science knowledge to select, transform, or extract characteristics, properties, and attributes from raw data **Feature extraction**: A type of feature engineering that involves taking multiple features to create a new one that would improve the accuracy of the algorithm **Feature selection**: A type of feature engineering that involves select the features in the data that contribute the most to predicting the response variable **Feature transformation**: A type of feature engineering that involves modify existing features in a way that improves accuracy when training the model

**Filtering**: The process of selecting a smaller part of a dataset based on specified values and using it for viewing or analysis

**First-party data**: Data that was gathered from inside your own organization **Float**: A number that contains a decimal

**FOR loop**: A piece of code that iterates over a sequence of values **Format()**: A function that formats specific values and inserts them inside the string’s placeholder

**Forward selection**: A stepwise variable selection process that begins with the null mode—with zero independent variables—considers all possible variables to add; incorporates the independent variable that contributes the most explanatory power to the model **Function**: A group of related statements that perform a specific task

## G

**Generator()**: A function that returns an object (iterator) which can be iterated over (one value at a time)

**Global outliers**: Values that are completely different from the overall data group and have no association with any other outliers

**Global variable**: A variable that can be accessed from anywhere in a program or script **Gradient boosting machines (GBMs)**: Model ensembles that use gradient boosting

**Gradient boosting**: A boosting methodology where each base learner in the sequence is built to predict the residual errors of the model that preceded it **GridSearch**: A tool to confirm that a model achieves its intended purpose by systematically checking every combination of hyperparameters to identify which set produces the best results, based on the selected metric

**Groupby()**: A method that splits objects into categories, applies a function, and combines the results

**Grouping**: The process of aggregating individual observations of a variable into groups

## H

**Hackathon**: An event where programmers and data professionals come together and work on a project

**Head()**: A function that returns a preview of the column names and the first few rows of a dataset

**Heatmap**: A type of data visualization that depicts the magnitude of an instance or set of values based on two colors

**Help()**: A Python help function used to display the documentation of modules, functions, classes, keywords, and more

**Histogram**: A data visualization that depicts an approximate representation of the distribution of values in a dataset

**Hold-out sample**: A random sample of observed data that is not used to fit the model

**Homoscedasticity assumption**: An assumption of simple linear regression stating that the variation of the residuals (errors) is constant or similar across the model

**Hyperparameters**: Parameters that can be set by the modeler before the model is trained **Hyperparameter tuning**: Refers to changing parameters that directly affect how the model trains, before the learning process begins

**Hypothesis**: A theory or an explanation, based on evidence, that is not yet proven true

**Hypothesis testing**: A statistical procedure that uses sample data to evaluate an assumption about a population parameter

## I

**IF statement**: A reserved keyword that sets up a condition in Python

**Immutability**: Refers to a data structure or element’s internal state cannot be changed **Immutable data type:** A data type for which the values can never be altered or updated **Implicit conversion**: Refers to when Python automatically converts one data type to another without user involvement

**Independent events**: The concept that two events are independent if the occurrence of one event does not change the probability of the other event

**Independent observation assumption**: An assumption of simple linear regression stating that each observation in the dataset is independent

**Independent variable (X)**: The variable whose trends are associated with the dependent variable

**Index**: A numbered position of each element in an ordered list **Inertia**: The sum of the squared distances between each observation and its nearest centroid **Inferential statistics**: A type of statistics that uses sample data to draw conclusions about a larger population

**Info()**: Gives the total number of entries, along with the data types—called Dtypes in pandas—of the individual entries

**Inheritance**: Refers to letting a programmer build relationships between concepts and group them together to reduce code duplication

**Input validation**: The practice of thoroughly analyzing and double-checking to make sure data is complete, error-free, and high-quality

**Input**: Information entered into a program

**Input()**: A Python function that can be used to ask a question in a message and store the answer in a variable

**Insert()**: A function that takes an index as the first parameter and an element as the second parameter

**Instance variable**: A variable that is declared in a class outside of other methods or blocks **Instantiation**: Refers to creating a copy of the class that inherits all class variables and methods

**Int64**: A standard integer data type, representing numbers somewhere between negative nine quintillion and positive nine quintillion

**Integer**: A data type used to represent whole numbers without fractions **Integrated Development Environment (IDE)**: A piece of software that has an interface to write, run, and test a piece of code

**Interaction term**: Represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

**Intercept (constant 𝐵**0**)**: The y value of the point on the regression line where it intersects with the y-axis

**Interpersonal skills**: Traits that focus on communicating and building relationships

**Interquartile range**: The distance between the first quartile (Q1) and the third quartile (Q3)

**Intersection()**: A method that finds the elements that two sets have in common

**Interval**: A sample statistic plus or minus the margin of error

**Interval estimate**: A calculation that uses a range of values to estimate a population parameter

**Is**: A rule that checks objects and classes for ancestry

**Items()**: A dictionary method to return key value pairs in a dictionary **Iterable**: An object that can be looped over, or iterated over

**Iteration**: The repeated execution of a set of instructions, where one iteration—or loop—is the single execution of a block of code

## J

**Joining**: The process of augmenting data by adding values from other datasets; one of the six practices of EDA

**JSON file**: A data storage file that is saved in a JavaScript format **Jupyter Notebook**: An open-source web application for creating and sharing documents containing live code, mathematical formulas, visualizations, and text

## K

**K-means**: An unsupervised partitioning algorithm used to organize unlabeled data into groups, or clusters

**Kernel**: An underlying core program, like Python

**Keys()**: A dictionary method to only retrieve the keys in a dictionary

**Keyword**: A predefined and reserved word in Python that has a special meaning and is used to define the syntax

## L

**Label encoding**: Data transformation technique where each category is assigned a unique number instead of a qualitative value

**Leaf node**: The nodes where a final prediction is made

**learning\_rate**: In XGBoost, a hyperparameter that specifies how much weight is given to each consecutive tree’s prediction in the final ensemble

**Len()**: A function used to measure the length of strings

**Library**: A reusable chunk of code. In Python, a collection of related packages and modules, and documentation

**Likelihood**: The probability of observing the actual data, given some set of beta parameters

**Line**: A collection of an infinite number of points extending in two opposite directions

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Linearity assumption**: An assumption of simple linear regression stating that each predictor variable (Xi) is linearly related to the outcome variable (Y)

**Link function**: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Linkage**: The method used to determine which points/clusters to merge

**List comprehension**: Refers to writing functions in just one line to create a new list based on sequences or a range

**List**: A data structure that helps store and manipulate an ordered collection of items

**Literacy rate**: The percentage of the population in a given age group that can read and write

**Log-Odds function**: (Refer to **logit**)

**Logical operator**: An operator that returns a logical data type

**Logistic regression**: A technique that models a categorical dependent variable (Y) based on one or more independent variables (X)

**Logit**: The logarithm of the odds of a given probability

**Loss function**: A function that measures the distance between the observed values and the model’s estimated values

**Lower limit**: When constructing an interval, the calculation of the sample means minus the margin of error

## M

**Machine learning**: The use and development of algorithms and statistical models to teach computer systems to analyze and discover patterns in data

**MAE (Mean Absolute Error)**: The average of the absolute difference between the predicted and actual values

**Magic commands**: Commands that are built into IPython to simplify common tasks

**Magics**: (Refer to **magic commands**)

**MANCOVA (Multivariate Analysis of Covariance)**: An extension of ANCOVA and MANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables, while controlling for covariates

**MANOVA (Multivariate Analysis of Variance)**: An extension of ANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables

**Margin of error**: The maximum expected difference between a population parameter and a sample estimate

**max\_depth**: In tree-based models, a hyperparameter that controls how deep each base learner tree will grow

**max\_features**: In decision tree and random forest models, a hyperparameter that specifies the number of features that each tree randomly selects during training called “colsample\_bytree” in XGBoost  **Maximum Likelihood Estimation (MLE)**: A technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data **Mean**: The average value in a dataset

**Measure of central tendency**: A value that represents the center of a dataset

**Measure of dispersion**: A value that represents the spread of a dataset, or the amount of variation in data points

**Measure of position**: A method by which the position of a value in relation to other values in a dataset is determined

**Measures**: Numeric values that can be aggregated or placed in calculations

**Median**: The middle value in a dataset

**Mentor**: Someone who shares knowledge, skills, and experience to help another grow both professionally and personally

**Merging**: A method to combine two (or more) different data frames along a specified starting column(s)

**Method**: A function that defines and performs behaviors like computation

**Metrics**: Methods and criteria used to evaluate data

**min\_child\_weight**: In XGBoost models, a hyperparameter indicating that a tree will not split a node if it results in any child node with less weight than this value called “min\_samples\_leaf” in decision tree and random forest models

**min\_samples**: In DBSCAN clustering models, a hyperparameter that specifies the number of samples in an ε-neighborhood for a point to be considered a core point (including itself)

**min\_samples\_leaf**: In decision tree and random forest models, a hyperparameter that defines the minimum number of samples for a leaf node called “min\_child\_weight” in XGBoost

**min\_samples\_split**: In decision tree and random forest models, a hyperparameter that defines the minimum number of samples that a node must have to split into more nodes

**Missing data**: A data value that is not stored for a variable in the observation of interest

**Mode**: The most frequently occurring value in a dataset

**Model assumptions**: Statements about the data that must be true in order to justify the use of a particular modeling technique

**Model selection**: The process of determining which model should be the final product and put into production

**Model validation**: The set of processes and activities intended to verify that models are performing as expected

**Modularity**: The ability to write code in separate components that work together and that can be reused for other programs

**Module**: A simple Python file containing a collection of functions and global variables **MSE (Mean Squared Error)**: The average of the squared difference between the predicted and actual values

**Multiple linear regression**: A technique that estimates the relationship between one continuous dependent variable and two or more independent variables

**Multiple regression**: (Refer to multiple linear regression)

**Multiplication rule (for independent events)**: The concept that if the events A and B are independent, then the probability of both A and B happening is the probability of A multiplied by the probability of B

**Mutability**: The ability to change the internal state of a data structure

**Mutually exclusive**: The concept that two events are mutually exclusive if they cannot occur at the same time

## N

**n\_clusters**: In K-means and agglomerative clustering models, a hyperparameter that specifies the number of clusters in the final model

**n\_estimators**: In random forest and XGBoost models, a hyperparameter that specifies the number of trees your model will build in its ensemble

**Naive Bayes**: A supervised classification technique that is based on Bayes’s Theorem with an assumption of independence among predictors

**Naming conventions**: Consistent guidelines that describe the content, creation date, and version of a file in its name

**Naming restrictions**: Rules built into the syntax of the language itself that must be followed **Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa **Nested loop**: One loop inside of another

**No multicollinearity assumption**: An assumption of simple linear regression stating that no two independent variables (Xi and Xj) can be highly correlated with each other

**Non-null count**: The total number of data entries for a data column that are not blank **Non-probability sampling**: A sampling method that is based on convenience or the personal preferences of the researcher, rather than random selection

**None**: A special data type in Python used to indicate that things are empty or that they return nothing

**Nonprofit**: A group organized for purposes other than generating profit; often aims to further a social cause or provide a benefit to the public

**Nonresponse bias**: Refers to when certain groups of people are less likely to provide responses

**Normal distribution**: A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped

**Normality assumption**: An assumption of simple linear regression stating that the residuals are normally distributed

## O

**Object**: A collection of data that consists of variables and methods or functions

**Object type**: A component category, usually associated with its respective class

**Object-oriented programming**: Programming modeled around data objects as opposed to functions

**Objective probability**: A type of probability based on statistics, experiments, and mathematical measurements

**Observed values**: The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

**One hot encoding**: A data transformation technique that turns one categorical variable into several binary variables

**One-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of one categorical variable

**Open data**: Data that is available to the public and free to use, with guidance on how to navigate the datasets and acknowledge the source

**Ordinary least squares estimation (OLS)**: A common way to calculate linear regression coefficients

**Outcome variable (Y)**: (Refer to dependent variable)

**Outliers**: Observations that are an abnormal distance from other values or an overall pattern in a data population

**Output**: A message stating what to do next

**Overfitting**: When a model fits the observed or training data too specifically and is unable to generate suitable estimates for the general population

## P

**P-value**: The probability of observing results as extreme as those observed when the null hypothesis is true

**PACE**: A workflow data professionals can use to remain focused on the end goal of any given dataset; stands for plan, analyze, construct, and execute

**Package**: A fundamental unit of shareable code that others have developed for a specific purpose

**Pandas**: A Python package used for data processing

**Parameter**: A characteristic of a population

**Percentile**: The value below which a percentage of data falls

**Personally identifiable information (PII)**: Information that permits the identity of an individual to be inferred by either direct or indirect means.

**Plan stage**: Stage of the PACE workflow where the scope of a project is defined and the informational needs of the organization are identified

**Point estimate**: A calculation that uses a single value to estimate a population parameter **Poisson distribution**: A probability distribution that models the probability that a certain number of events will occur during a specific time period

**Pop()**: A method that extracts an element from a list by removing it at a given index

**Popularity bias**: The phenomenon of more popular items being recommended too frequently **Population**: Every possible element that a data professional is interested in measuring

**Population proportion**: The percentage of individuals or elements in a population that share a certain characteristic

**Positive correlation**: A relationship between two variables that tend to increase or decrease together.

**Post hoc test**: An ANOVA test that performs a pairwise comparison between all available groups while controlling for the error rate **Posterior probability**: The probability of an event occurring after taking into consideration new information

**Precision**: The proportion of positive predictions that were correct to all positive predictions

**Predicted values**: The estimated Y values for each X calculated by a model

**Predictor variable**: (Refer to independent variable)

**Presenting**: The process of making a cleaned dataset available to others for analysis or further modeling; one of the six practices of EDA

**Prior probability**: Refers to the probability of an event before new data is collected

**Probability**: The branch of mathematics that deals with measuring and quantifying uncertainty

**Probability distribution**: A function that describes the likelihood of the possible outcomes of a random event

**Probability sampling**: A sampling method that uses random selection to generate a sample

**Program**: A series of instructions written so that a computer can perform a certain task, independent of any other application

**Purposive sample**: A method of non-probability sampling that involves researchers selecting participants based on the purpose of their study

**Python**: A general-purpose programming language

## Q

**Quartile**: A value that divides a dataset into four equal parts

## 

## R

**R2 (The Coefficient of Determination)**: Measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X

**RACI chart:** A visual that helps to define roles and responsibilities for individuals or teams to ensure work gets done efficiently; lists who is responsible, accountable, consulted, and informed for project tasks

**Random experiment**: A process whose outcome cannot be predicted with certainty

**Random forest**: An ensemble of decision trees trained on bootstrapped data with randomly selected features

**Random seed**: A starting point for generating random numbers **Random variable**: A variable that represents the values for the possible outcomes of a random event

**Range**: The difference between the largest and smallest value in a dataset **Range()**: A function that returns a sequence of numbers starting from zero, increments by one by default, and stops before the given number

**Recall**: The proportion of actual positives that were identified correctly to all actual positives

**Recommendation systems**: Unsupervised learning techniques that use unlabeled data to offer relevant suggestions to users

**Refactoring**: The process of restructuring code while maintaining its original functionality **Regression analysis**: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables **Regression coefficient**: The estimated betas in a regression model **Regression models**: (Refer to **regression analysis**)

**Regularization**: A set of regression techniques that shrinks regression coefficient estimates towards zero, adding in bias, to reduce variance

**Remove()**: A method that removes an element from a list

**Representative sample**: A sample that accurately reflects the characteristics of a population

**Reshape**: A NumPy method that gives a new shape to an array without changing its data **Residual**: The difference between observed or actual values and the predicted values of the regression line

**Response variable**: (Refer to **dependent variable**)

**Return**: A reserved keyword in Python that makes a function do work to produce new results **Reusability**: Refers to defining code once and using it many times without having to rewrite it **Root node**: The first node of the tree, where the first decision is made

## 

## S

**Sample**: A segment of a population that is representative of the entire population

**Sample size**: The number of individuals or items chosen for a study or experiment

**Sample space**: The set of all possible values for a random variable

**Sampling**: The process of selecting a subset of data from a population

**Sampling bias**: Refers to when a sample is not representative of the population as a whole

**Sampling distribution**: A probability distribution of a sample statistic

**Sampling frame**: A list of all the items in a target population

**Sampling variability**: Refers to how much an estimate varies between samples

**Sampling with replacement**: Refers to when a population element can be selected more than one time

**Sampling without replacement**: Refers to when a population element can be selected only one time

**Simple random sample**: A probability sampling method in which every member of a population is selected randomly and has an equal chance of being chosen

**Scatterplot matrix**: A series of scatterplots that show the relationships between pairs of variables

**Script**: A collection of commands in a file designed to be executed like a program

**Second-party data**: Data that was gathered outside your organization but directly from the original source

**Self**: A parameter passed to a method or attributes used to instantiate an object

**Self-documenting code**: Code written in a way that is readable and makes its purpose clear

**Semantics**: Refers to the variables and objects that give meaning to Python code

**Sequence**: A positionally-ordered collection of items

**Series**: A one-dimensional labeled array capable of holding any data type

**Set**: A data structure in Python that contains only unordered, non-interchangeable elements; a Tableau term for a custom field of data created from a larger dataset based on custom conditions

**Set()**: A function that takes an iterable as an argument and returns a new set object

**Shrinkage**: (Refer to **learning\_rate**)

**Silhouette analysis**: The comparison of different models’ silhouette scores **Silhouette score**: The mean of the silhouette coefficients of all the observations in a model **Simple linear regression:** A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y **Single**: The minimum pairwise distance between clusters

**Slicing**: A method for breaking information down into smaller parts to facilitate efficient examination and analysis from different viewpoints

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Snowball sample**: A method of non-probability sampling that involves researchers recruiting initial participants to be in a study and then asking them to recruit other people to participate in the study

**Sorting**: The process of arranging data into a meaningful order for analysis

**Standard deviation**: A statistic that calculates the typical distance of a data point from the mean of a dataset

**Standard error**: The standard deviation of a sample statistic

**Standard error of the mean**: The sample standard deviation divided by the square root of the sample size

**Standard error of the proportion**: The square root of the sample proportion times one minus the sample proportion divided by the sample size **Standardization**: The process of putting different variables on the same scale

**Statistic**: A characteristic of a sample

**Statistical significance**: The claim that the results of a test or experiment are not explainable by chance alone

**Statistics**: The study of the collection, analysis, and interpretation of data

**Story**: A Tableau term for a group of dashboards or worksheets assembled into a presentation

**Stratified random sample**: A probability sampling method that divides a population into groups and randomly selects some members from each group to be in the sample

**Subjective probability**: A type of probability based on personal feelings, experience, or judgment

**String**: A sequence of characters and punctuation that contains textual information

**String literal**: A programming string used in code in which characters exist as the value themselves, rather than as variables

**String slice**: The portion of a string that can contain more than one character, also sometimes called a substring

**Structuring**: The process of taking raw data and organizing or transforming it to be more easily visualized, explained, or modeled; one of the six practices of EDA

**Sum of squared residuals (SSR)**: The sum of the squared difference between each observed value and its associated predicted value

**Summary statistics**: A method that summarizes data using a single number

**Supervised machine learning**: A category of machine learning that uses labeled datasets to train algorithms to classify or predict outcomes

**Supervised model**: A machine learning model that is used to make predictions about unseen events

**Symmetric difference()**: A method that finds elements from both sets that are mutually not present in the other

**Syntax**: The predetermined structure of a language that includes all required words, symbols, and punctuation, as well as their proper placement

**Systematic random sample**: A probability sampling method that puts every member of a population into an ordered sequence, chooses a random starting point in the sequence, and selects members for the sample at regular intervals

## 

## T

**Tableau**: A business intelligence and analytics platform that helps people visualize, understand, and make decisions with data

**Target population**: The complete set of elements that someone is interested in knowing more about

**Third-party data**: Data gathered outside your organization and aggregated

**Tolist()**: A NumPy method to convert arrays into lists

**Tree-based learning**: A type of supervised machine learning that performs classification and regression tasks

**Tuple**: A sequence containing elements of any data type

**Tuple()**: A function that transforms input into tuples

**Two-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of two categorical variables

**Type()**: A function used to identify a data type

## U

**Undercoverage bias**: Refers to when some members of a population are inadequately represented in a sample

**Union()**: A method that can be used to find all the elements from both sets **Unsupervised model**: A machine learning model that is used to discover the natural structure of the data, finding relationships within unlabeled data

**Upper limit**: When constructing an interval, the calculation of the sample means plus the margin of error

**Upsampling**: The process of taking observations from the minority class and either adding copies of those observations to the dataset or generating new observations to add to the dataset

## V

**Validating**: The process of verifying that the data is consistent and high quality; one of the six practices of EDA

**Values()**: A dictionary method that only retrieves the values from a dictionary

**Variable**: A value that refers to where information is stored

**Variable selection**: The process of determining which variables or features to include in a given model

**Variance inflation factors (VIF)**: Quantifies how correlated each independent variable is with all of the other independent variables

**Variance**: Refers to model flexibility and complexity, so the model learns from existing data; The average of the squared difference of each data point from the mean

**Vectorization**: A technique that uses functions instead of loops to iterate over code

**Voluntary response sample**: A method of non-probability sampling that consists of members of a population who volunteer to participate in a study

## 

## W

**Ward**: Merges two clusters whose merging will result in the lowest inertia **Weak learner**: A model that performs slightly better than randomly guessing **WHILE loop**: A loop that instructs your computer to continuously execute your code based on the value of a condition

## X

**XGBoost (extreme gradient boosting)**: An optimized GBM package

## Z

**Zero Frequency problem**: Occurs when the dataset has no occurrences of a class label and some value of a predictor variable together

**Z-score**: A measure of how many standard deviations below or above the population mean a data point is