Variables Glossary

Binarization	The process of transforming a single categorical variable into multiple binary variables, where each new binary variable is an "indicator" for one of the levels in the categorical variable. For example, the Binarization of the categorical variable Pet with levels "dog," "cat," and "gerbil" would result in three new binary variables: "Pet_dog," which has value "1" when the Pet variable has value "dog" and "0" otherwise; "Pet_cat," which has value "1" when the Pet variable has value "cat" and "0" otherwise; and "Pet_gerbil," which has value "1" when the Pet variable has value "gerbil" and "0" otherwise. This is also known as one-hot encoding.
Boolean/binary variables	Variables whose values are "true" or "false." Equivalently, these variables are those whose values are "1" or "0."
Categorical variables	Nonnumeric variables whose values take on one of a set of predefined values (e.g., state, gender, and ZIP Code). These are also known as factor variables in the R programming language.
Character variables	Nonnumeric variables. The terminology "character" or "string" is more often used to describe nonnumeric variables that are not considered to be categorical or factors. For example, variables storing an individual's address, or the comments made by an underwriter would be considered character variables.
Continuous variables	Numerical variables whose values can be anything on the real number line within the range associated with the variable. Sometimes, these are also referred to as "real" variables.
Date variables	A special type of numerical variable that represents dates (e.g., 1st of February, 2001). Like numerical variables, these variables have an order, but their interpretation is usually quite different, for example, subtracting dates makes sense but adding or multiplying them usually doesn't. Additionally, date variables have cyclical properties (e.g., weeks, months, and years).
Dimensionality	A characteristic of categorical variables, which refers to the number of levels of the variable. For example, the variable "State" might have dimensionality 50 (high), but the variable "Gender" might have dimensionality 2 (low).
Discrete variables	Numerical variables, which, unlike continuous variables, only take on specific values within the range (e.g., only integer values).

Factor	Nonnumeric variables whose values take on one of a set of predefined values (e.g., state, gender, and ZIP Code). These are also known as categorical variables, although "factor" is the terminology used in the R programming language. Note: Sometimes, "factor" can be used to refer to predictor variables, but for the purposes of this course, we will tend to reserve its use to describe the R variable type.
Feature	Another name for a variable, although there is a subtle difference, which will be covered in later modules. Generally, the terms are interchangeable.
Geospatial/location variables	Special types of variables that refer to locations in space, for example, latitude, longitude, ZIP Code, and so forth. These variables have additional properties to most other variable types, such as "distances" between values or with reference to other points of interest, such as "distance to nearest hospital."
Granularity	A characteristic of variables that refers to how precisely or to what level of detail a variable is measured. For example, you might record the location of a policyholder by his or her state (low granularity), ZIP Code, address, or GPS coordinates from his or her phone (most granular).
Level	One of the predefined values of a categorical or factor variable. For example, the categorical variable "Gender" might have the following levels: "Female," "Male," and "Other."
Numeric variables	Variables that have values that are numbers, usually with an associated range (e.g., Age, which will have values greater than 0).
Observation	A row in a dataset that represents the set of measurement/characteristic values for each variable for that particular row. For example, in a dataset containing information about a book of life insurance policies, an observation might be a single row in that dataset with all the information about that policy (i.e., the policyholder name, the sum insured, the age of the policy holder, etc.).
Order	A characteristic of variables that determines whether or not you can consider some values to be "greater" than other variables. If for every value of a variable you can say which of the other values are "greater" than it, and which of the other values are "less" than it, then the variable has an order that can be defined on it. For example, numerical variables have an order and some categorical variables can have an order defined on them ("Gold," "Silver," "Bronze," etc.).

Predictor variables	All variables used in a predictive model to predict the target variable, that is, all the variables used in a predictive model that are not the target variable.
Range	Lowest and highest values that a numerical variable can take. Note that there are other (more precise) definitions that say that the range refers to the set of possible values (so it would be different for continuous and discrete variables that have the same "end points"), but for the purposes of this course, we loosely refer to the range as the lowest and highest values (e.g., Age of a policyholder might have a range of 0 to 120).
Record	Another name for an observation and refers to a row in the dataset. The terms are interchangeable.
Target variables	The variable that we are trying to predict in a predictive model. For example, if we want to predict whether or not a policyholder will lapse his or her policy, then the variable representing "lapse"/"no-lapse" for each policy would be the target variable.
Time variables	Similar to date variables, but referring to the time of day, sometimes in addition to the date. These also have cyclical properties (e.g., minutes and hours).
Variables	A column in a dataset that represents a characteristic or measurement of the records in the data. For example, the "Age" and "Smoking Status" of individuals are considered variables.