# CSE303

Lecture 3: Exploratory Data Analysis

# ATTRIBUTES

- Data points or Samples are described by attributes.
- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
- Types
  - Nominal or Categorical
  - Ordinal
  - Binary
  - Numerical

# ATTRIBUTE TYPES

- Nominal: categories, states, or "names of things"
  - Hair color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- Ordinal: Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings
- Binary: Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important, e.g., gender
  - Asymmetric binary: outcomes not equally important. e.g., medical test (positive vs. negative)
- Numeric: represents quantity (integer or real-valued)
  - Temperature, length, counts, grade point, CGPA, salary etc.

# DISCRETE VS. CONTINUOUS ATTRIBUTES

- Discrete Attribute: has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute: has real numbers as attribute values
  - E.g., temperature, height, or weight
  - Continuous attributes are typically represented as floating-point variables

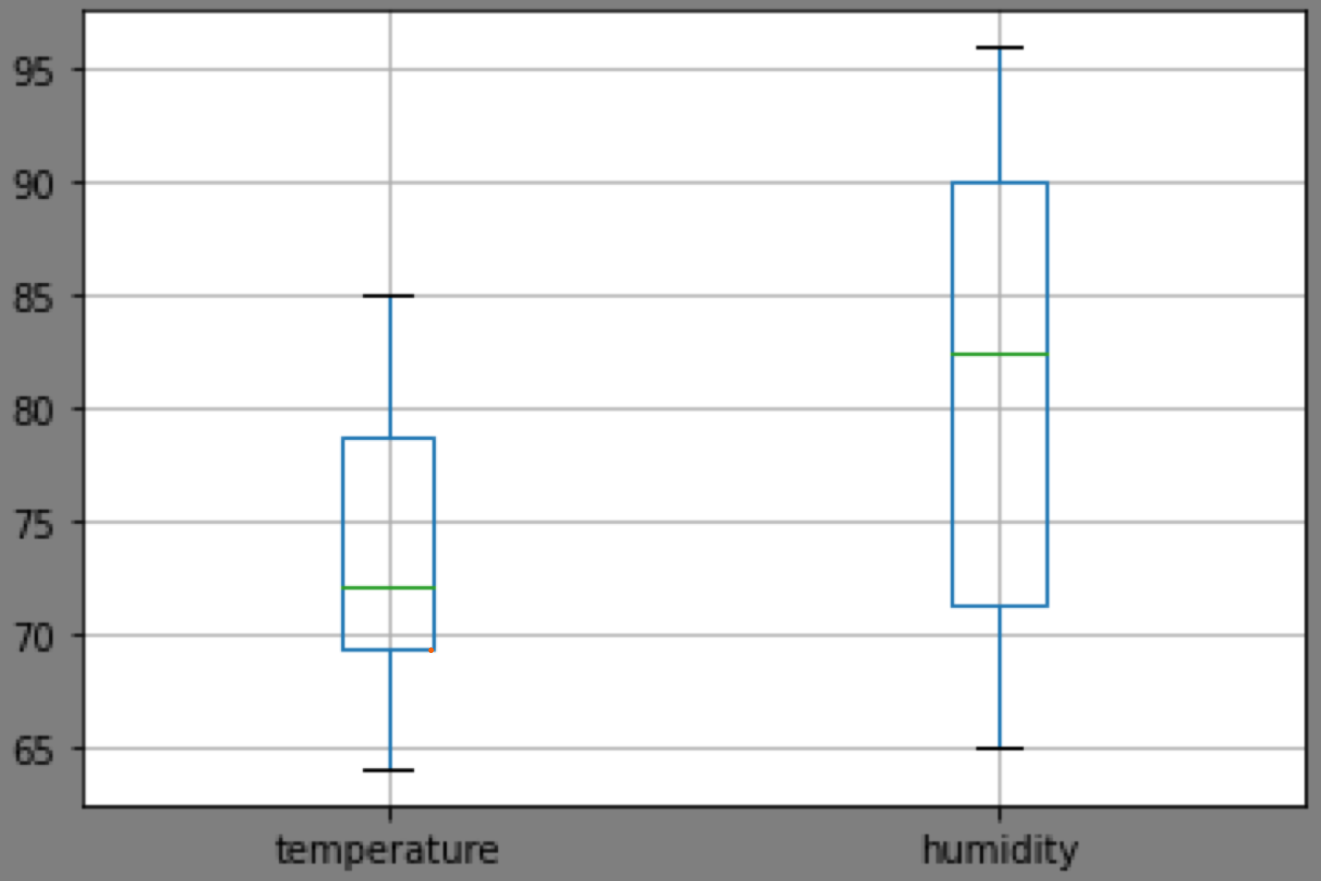
# A SAMPLE DATASET

| outlook | temperature | humidity | windy | play |
|---|---|---|---|---|
| sunny | 85 | 85 | FALSE | no |
| sunny | 80 | 90 | TRUE | no |
| overcast | 83 | 86 | FALSE | yes |
| rainy | 70 | 96 | FALSE | yes |
| rainy | 68 | 80 | FALSE | yes |
| rainy | 65 | 70 | TRUE | no |
| overcast | 64 | 65 | TRUE | yes |
| sunny | 72 | 95 | FALSE | no |
| sunny | 69 | 70 | FALSE | yes |
| rainy | 75 | 80 | FALSE | yes |
| sunny | 75 | 70 | TRUE | yes |
| overcast | 72 | 90 | TRUE | yes |
| overcast | 81 | 75 | FALSE | yes |
| rainy | 71 | 91 | TRUE | no |

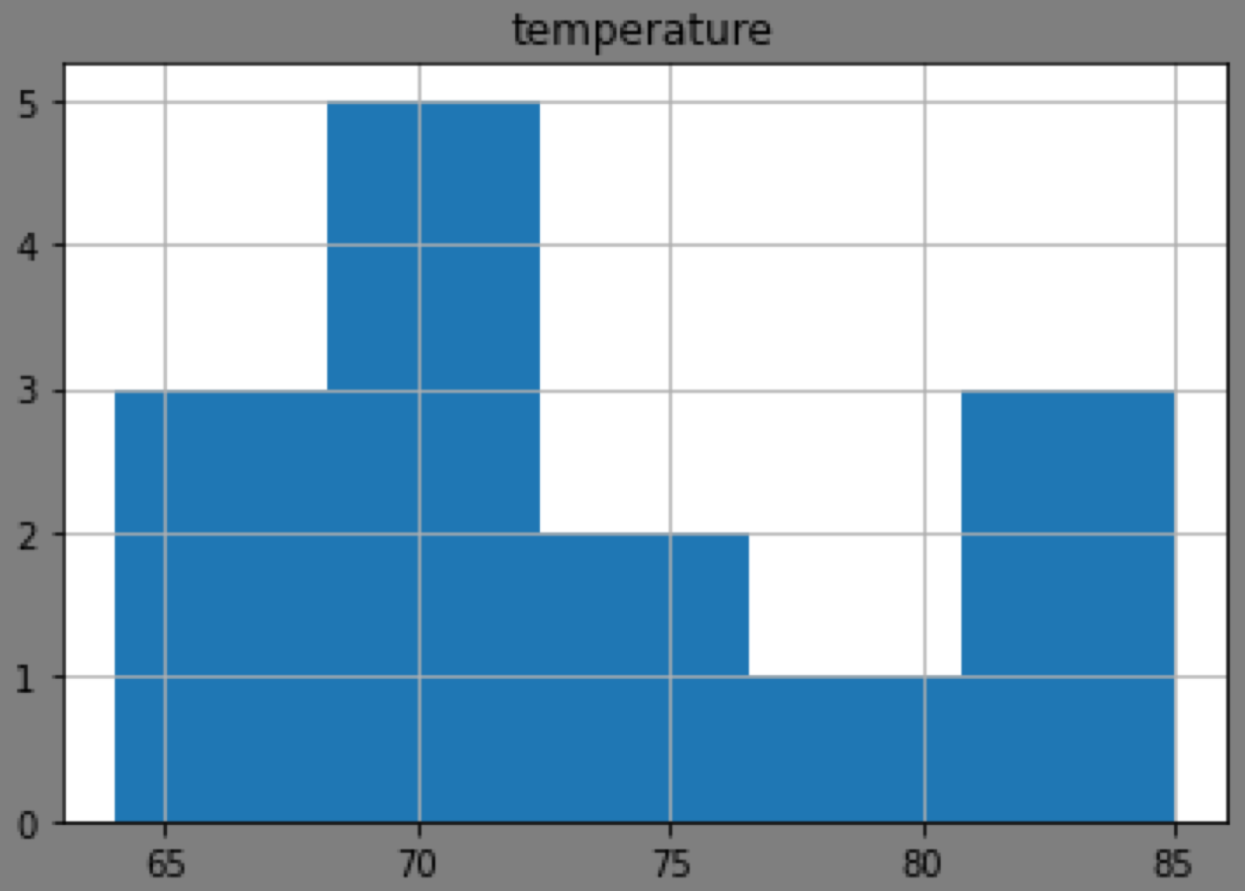
# EXPLORING DATA DISTRIBUTION

- There are many visual representation methods to explore the distribution of data.
  - Boxplot: a five-number summary (min, Q1, median, Q3, max)
  - Frequency Table: A tally of the count of numeric data values that fall into a set of intervals (bins).
  - Histogram: A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis.
  - Density Plot: smoothed version of Histogram.

# EXAMPLE: BOXPLOT



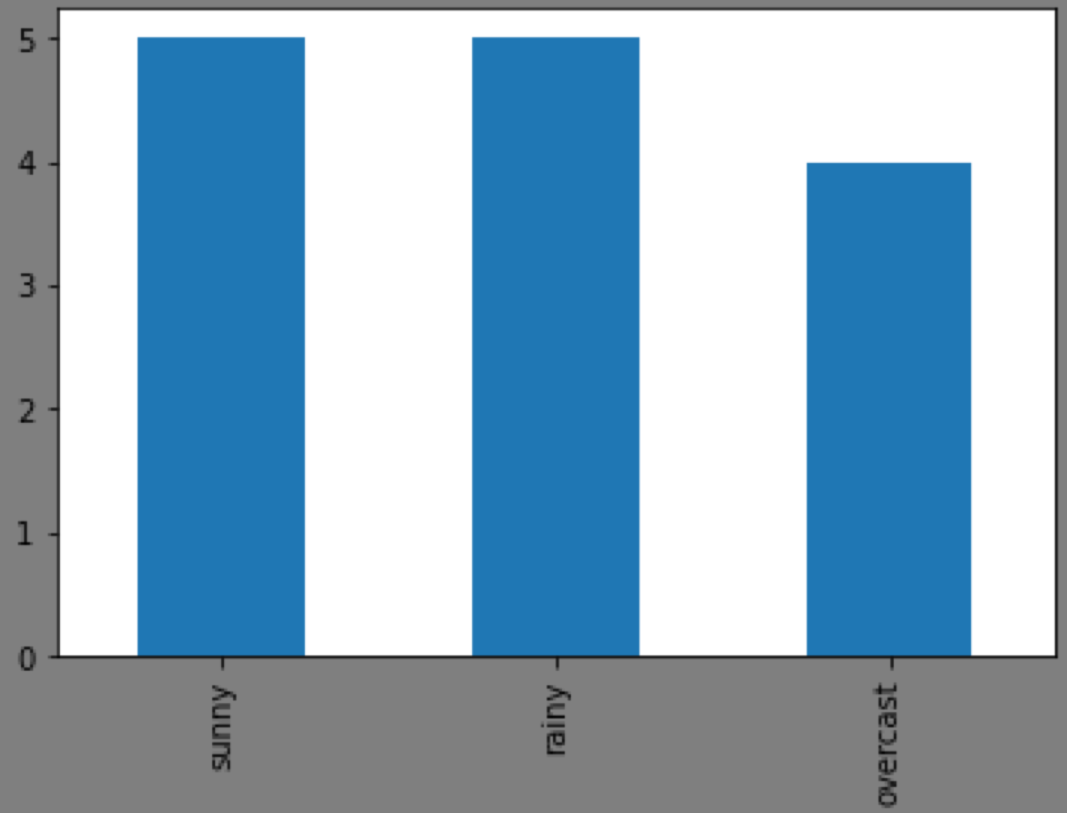
# EXAMPLE: HISTOGRAM



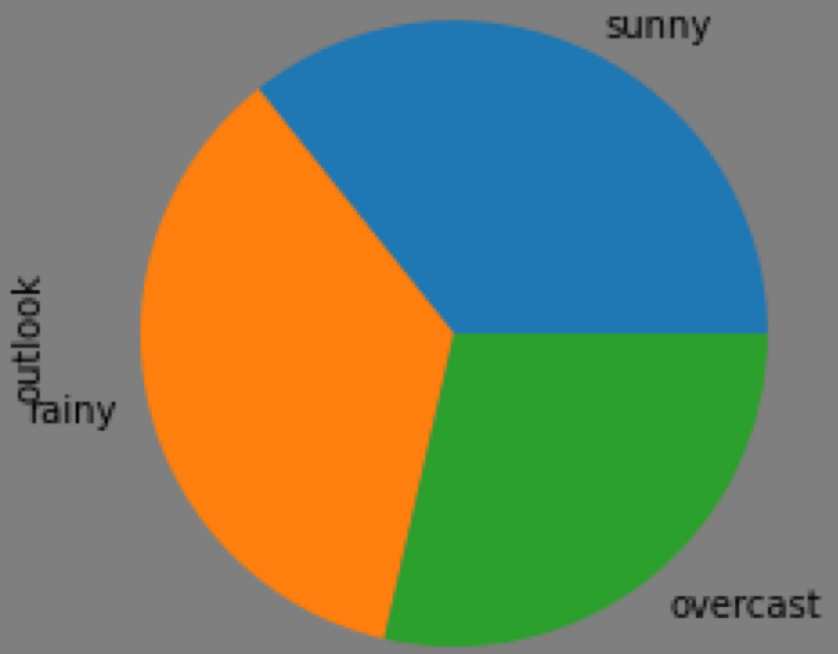
# EXPLORING NOMINAL AND BINARY DATA

- For nominal (categorical) data, simple proportions or percentages can give us the insight.
  - Mode: most commonly occurring category or value in a data set.
  - Expected value: When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.
  - Bar charts: The frequency or proportion of each category plotted as bars.
  - Pie charts: The frequency of proportion of each category plotted as wedges in a pie.

# EXAMPLE: BAR CHART



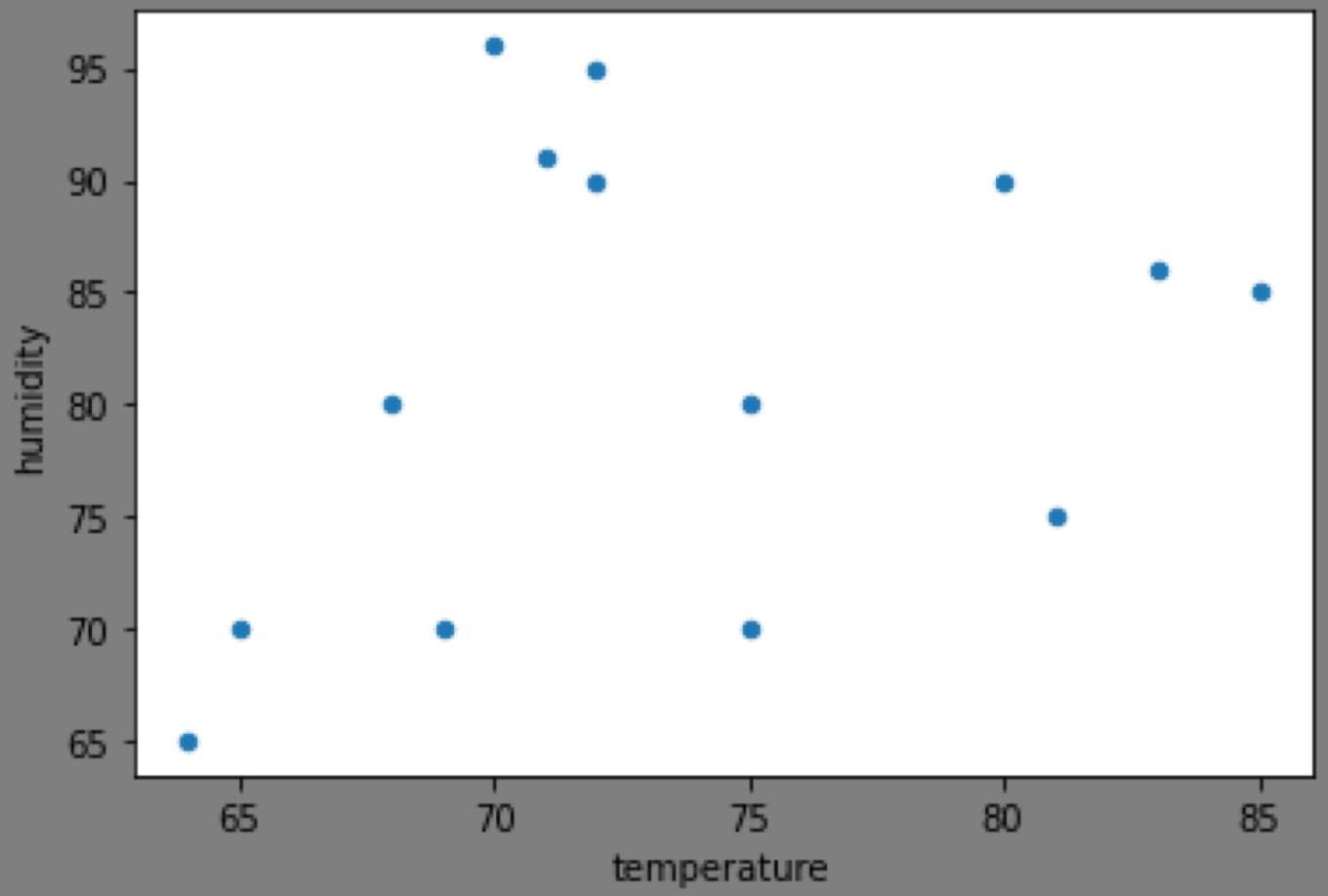
# EXAMPLE: PIE CHART



# EXPLORING TWO OR MORE VARIABLES

- Contingency Tables: A tally of counts between two or more categorical variables
- Scatterplots: shows relationship between two numeric variables. Not suitable for many data points.
- Hexagonal binning: A plot of two numeric variables with the records binned into hexagons
- Boxplots: A simple way to visually compare the distributions of a numeric variable grouped according to a categorical variable.

# EXAMPLE: SCATTER PLOT



# USEFUL RESOURCES

- Chapter 1, Practical Statistics for Data Scientists by Bruce and Bruce
- https://pandas.pydata.org/pandas-docs/stable/reference/index.html
- https://etav.github.io/python/count_basic_freq_plot.html

# THANK YOU