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

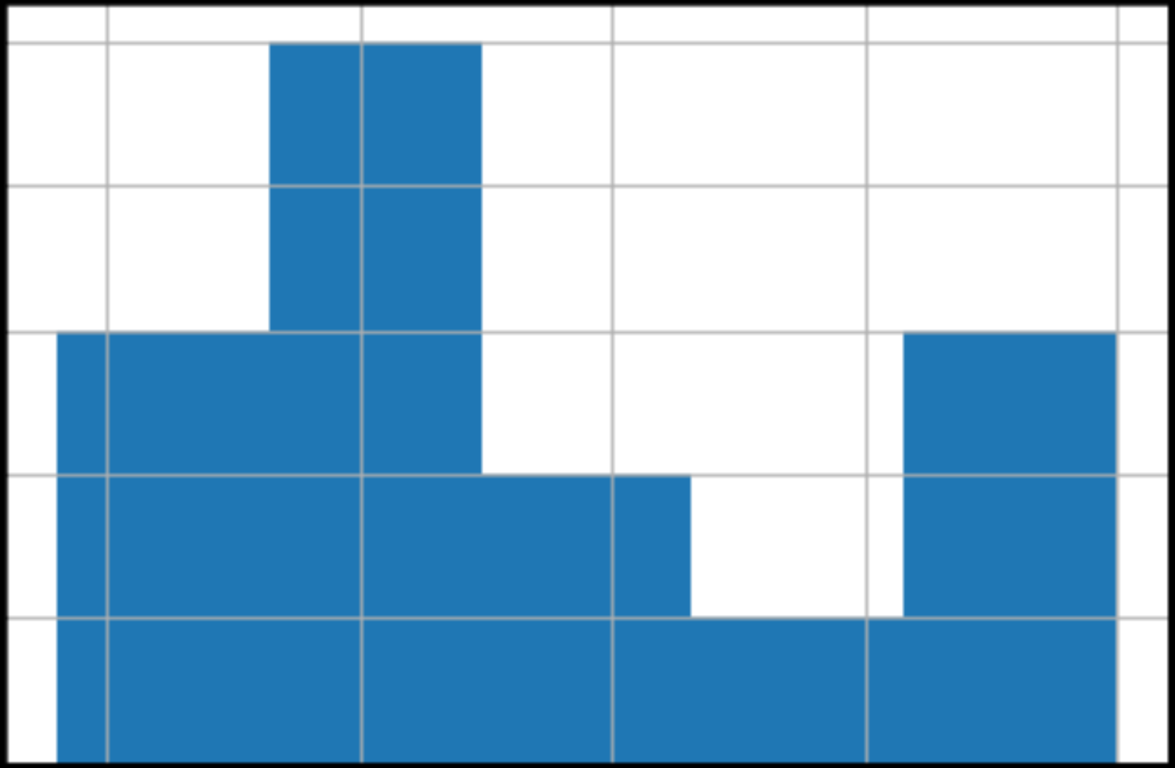
A sample Dataset

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

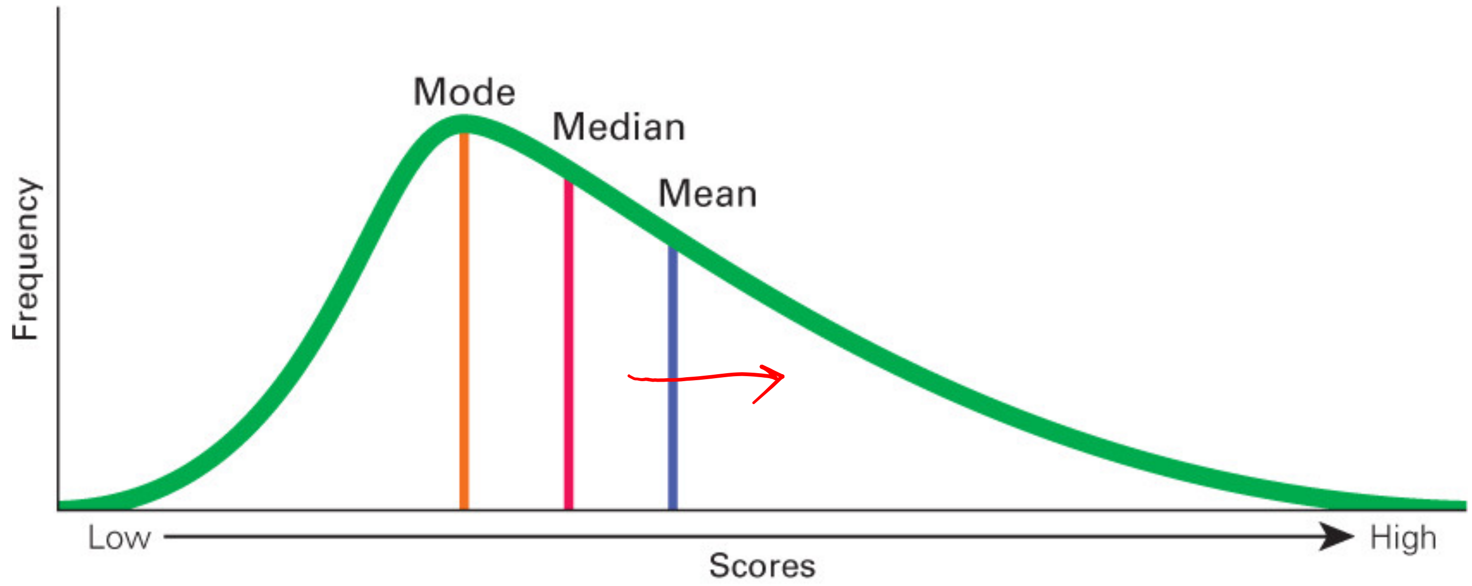
- There are many visual representation methods to explore the distribution of data.
  - Boxplot: a five-number summary (min, Q1, median, Q3, max)
  - Frequency Table: A tally of the count of numeric data values that fall into a set of intervals (bins).
  - Histogram: A plot of the frequency table with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

Example: HISTOGRAM



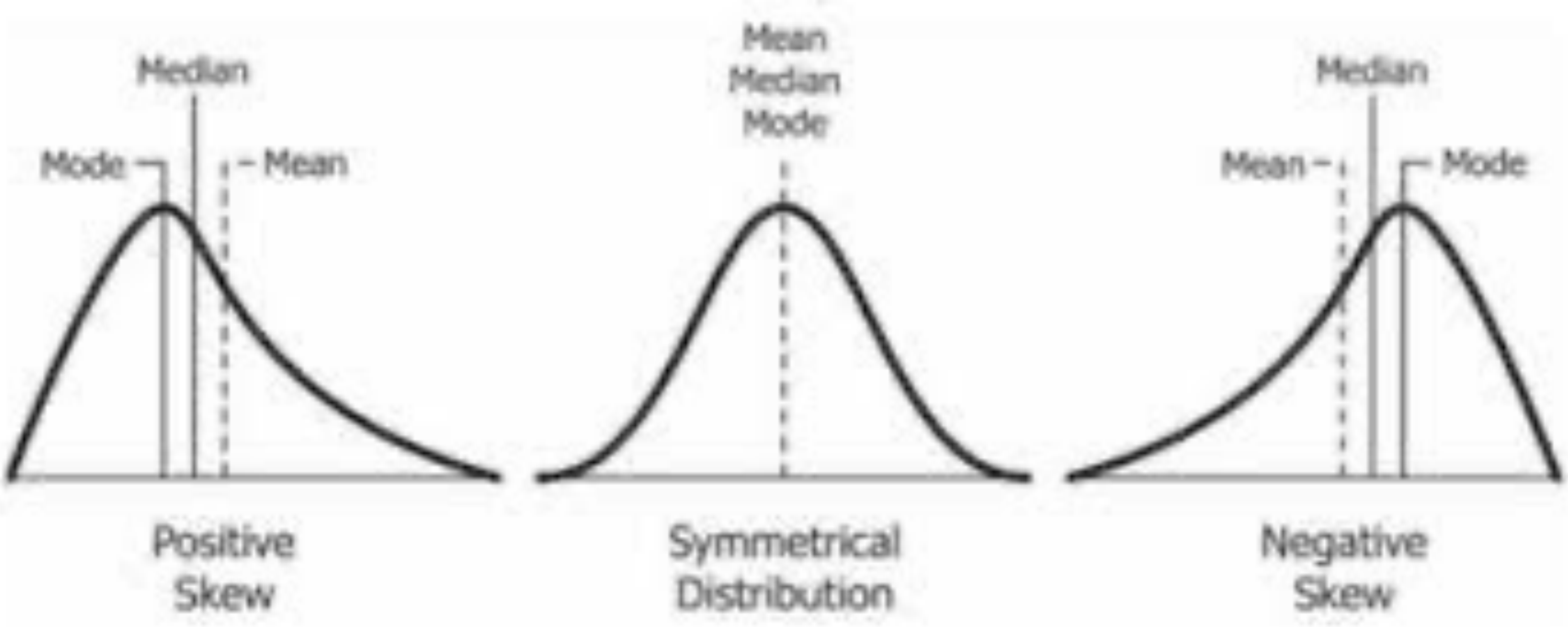
ANOTHER EXAMPLE of histogram

Runs Scored in First Innings	Frequency
80-100	3
100-120	4
120-140	7
140-160	12
160-180	3
180-200	3
200-220	2



(a) Right-skewed distribution

(b) Left-skewed distribution


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

DEFINITION OF BOXPLOT

- It is a 5-number summary
- MIN, LOWER Quartile (Q1), Median, Upper Quartile (Q3), MAX
- BOXPLOT is efficient to find outliers.
- Upper Extreme = Q3 + IQR X 1.5
- Lower Extreme = Q1 – IQR X 1.5
- If any data point exists that does not contained within the boundary of Lower and Upper Extreme then those datapoints can be identified as OUTLIERS.

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

64, 64, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85

Min = 64
Q1 = 25% X 14 = 3.5th value = 68.5
Median = 72
Q3 = 75% X 14 = 10.5th value = 77.5
Max = 85

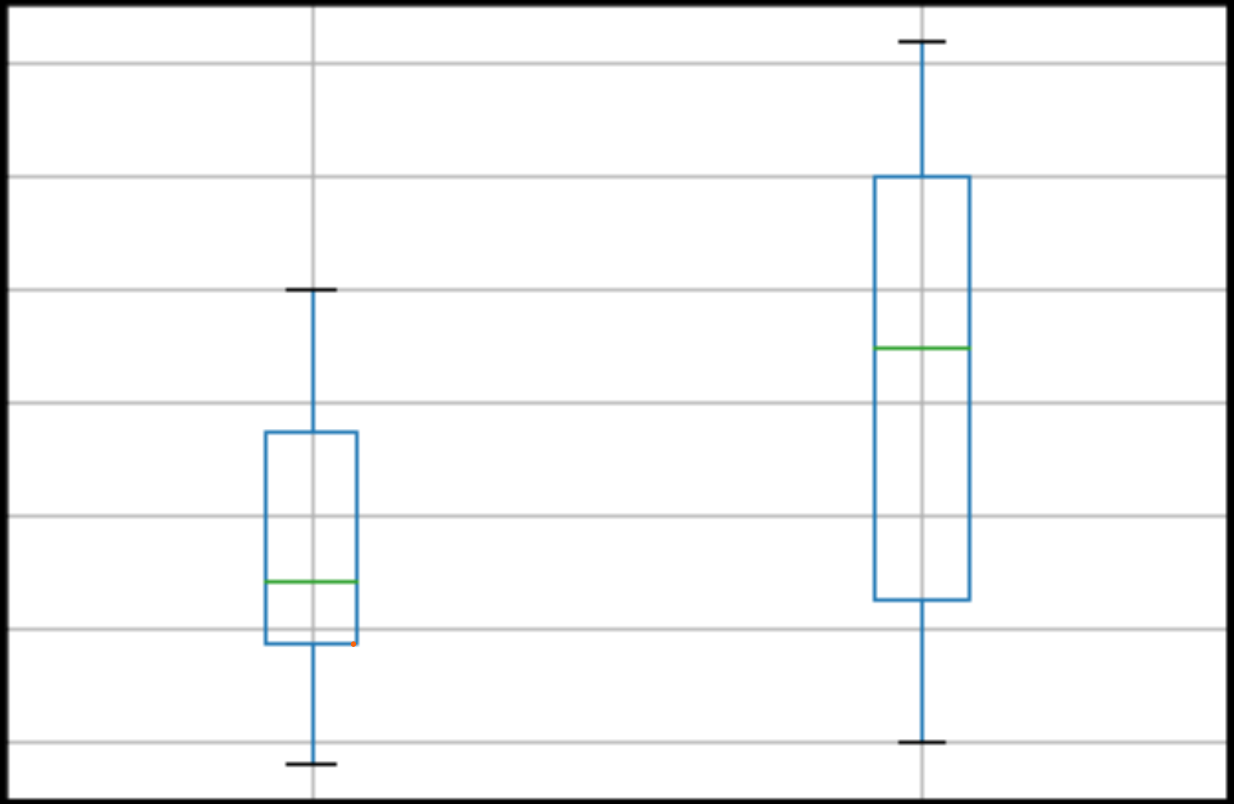
IQR = Q3-Q1 = 9
Upper Extreme = Q3 + IQR X 1.5
= 77.5 + 9 X 1.5
= 91

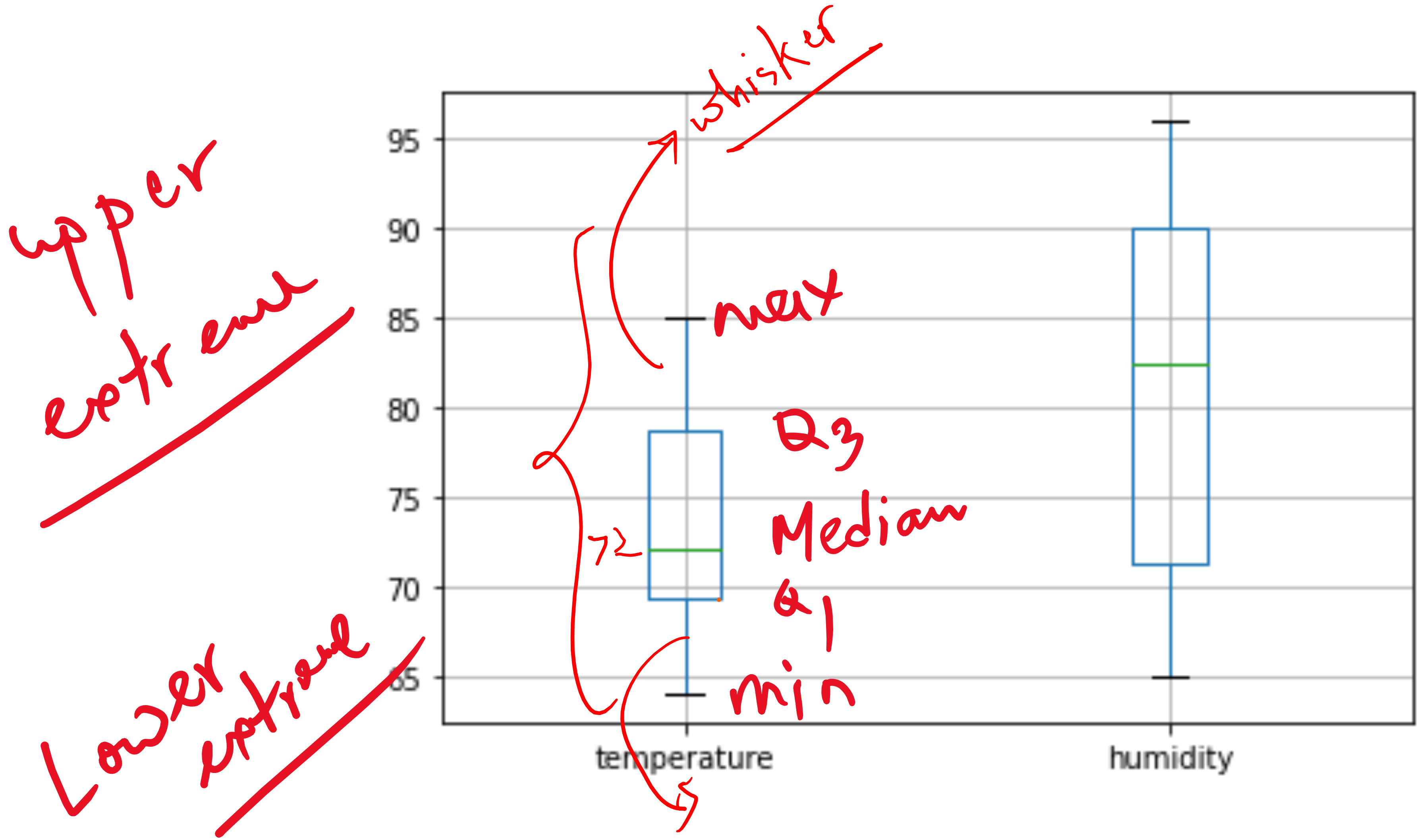
Lower Extreme = Q1 – IQR X 1.5
= 68.5 – 9 X 1.5
= 55

Any values greater than Upper Extreme or smaller than Lower Extreme would be called as OUTLIERS.

Upper Whisker – it will be extended till Max or Upper Extreme, whichever is lower.
Lower Whisker – it will be extended till Min or Lower Extreme, whichever is higher.
Each data point which are greater than Upper Extreme or smaller than Lower Extreme would be represented as a dot in the Boxplot.

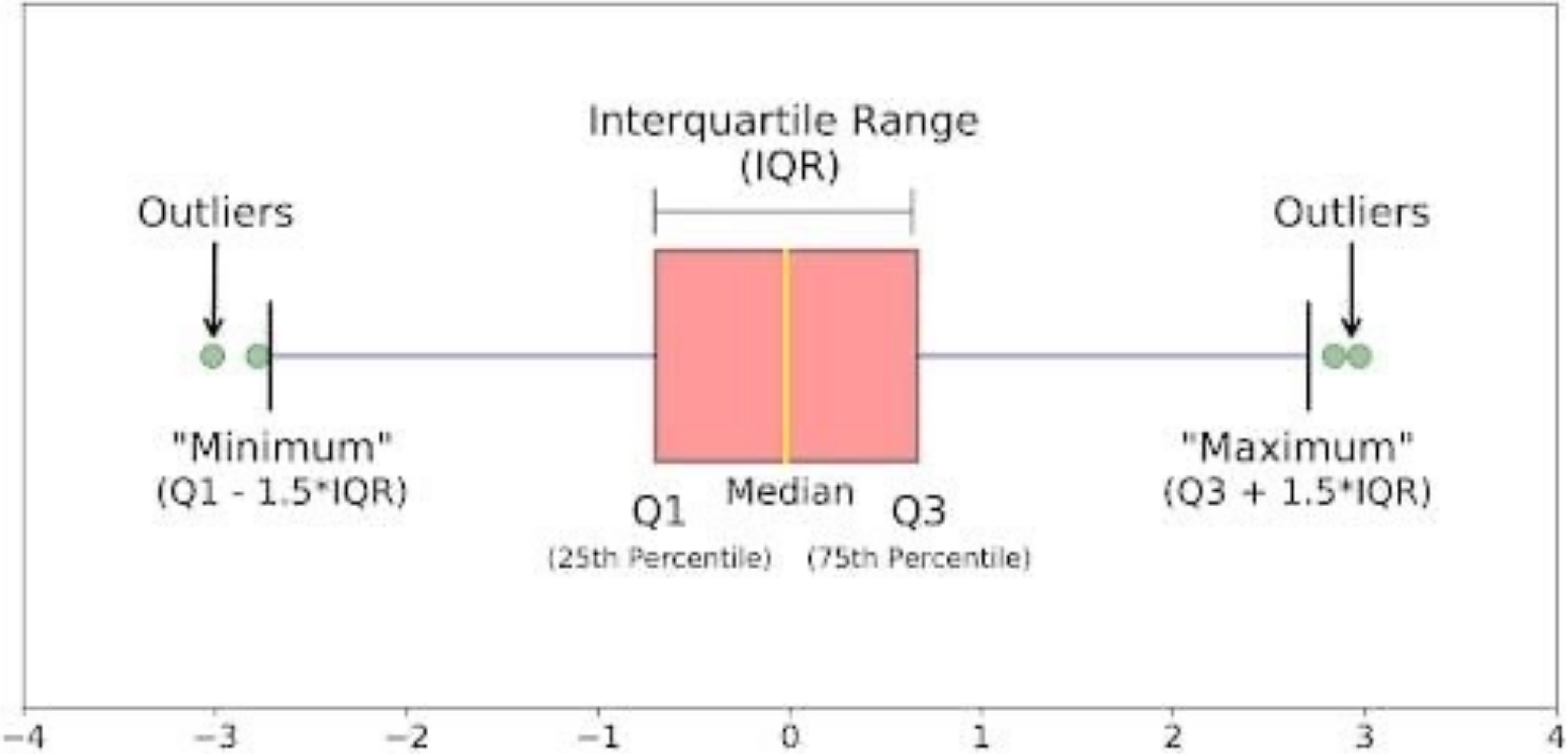
Example: Boxplot



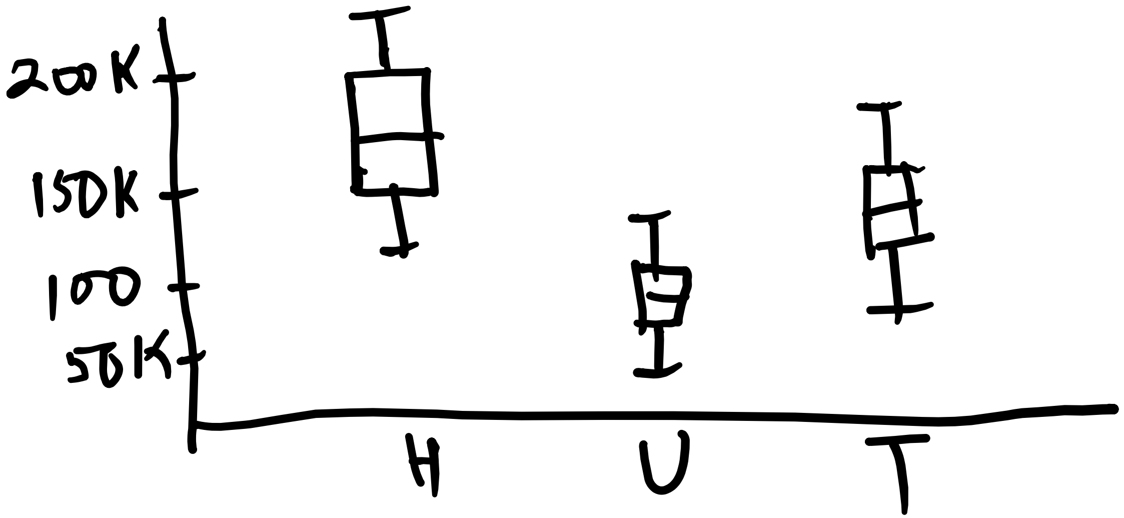


upper extreme
Lower extreme
whisker
max
Q3
72
Median
Q1
min
95
90
85
80
75
70
65
temperature
humidity

- For some distributions/data sets, you will find that you need more information than the measures of central tendency (median, mean and mode). You need to have information on the variability or dispersion of the data. A boxplot is a graph that gives you a good indication of how the values in the data are spread out. Although boxplots may seem primitive in comparison to a histogram or density plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or data sets.



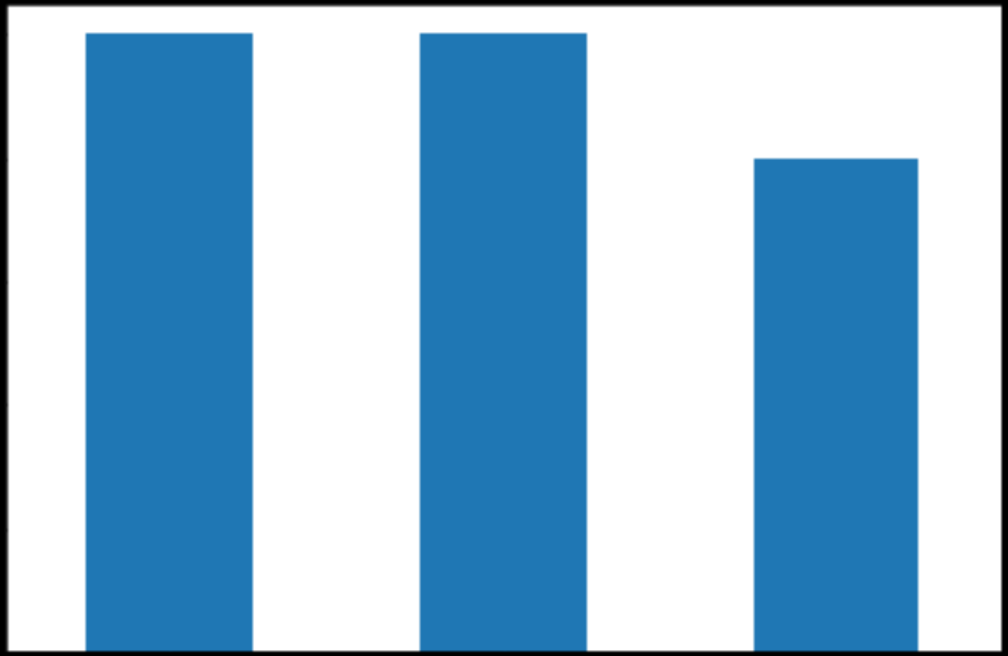
HouseType	Price
H	200000
H	150000
U	75000
U	54000
T	100000
T	90000

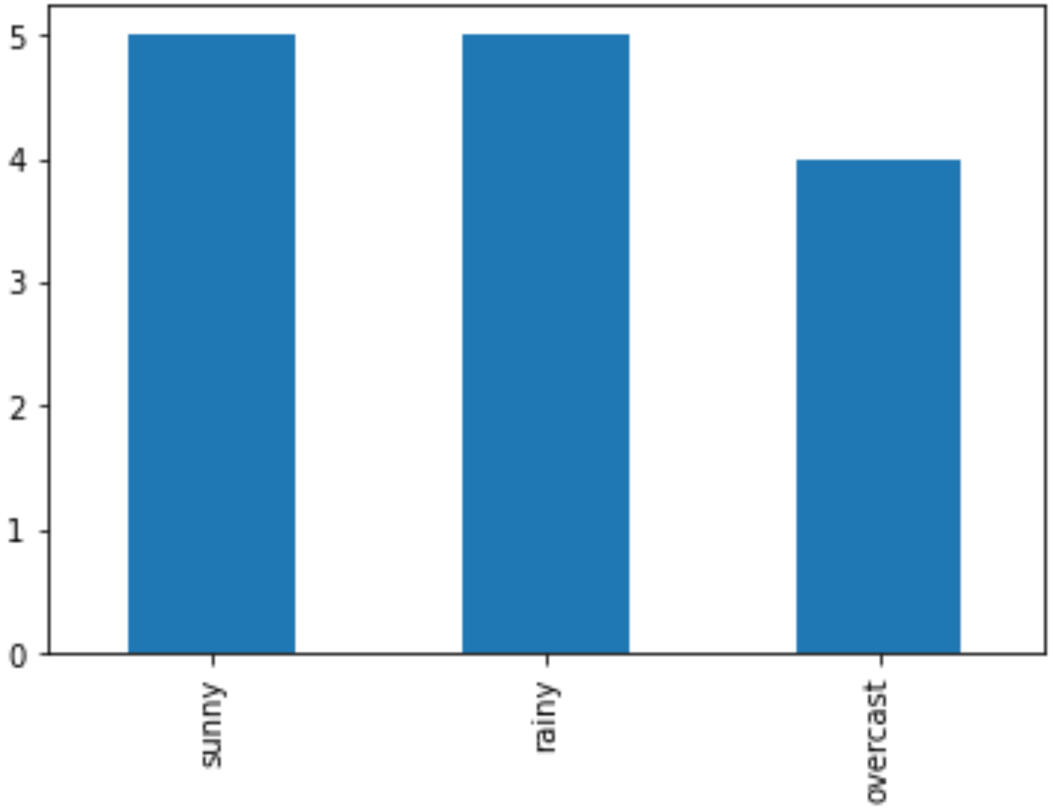


Exploring nominal and binary data

- 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.

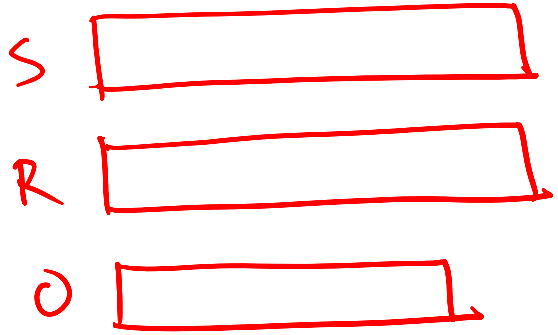
EXAMPLE: Bar CHART





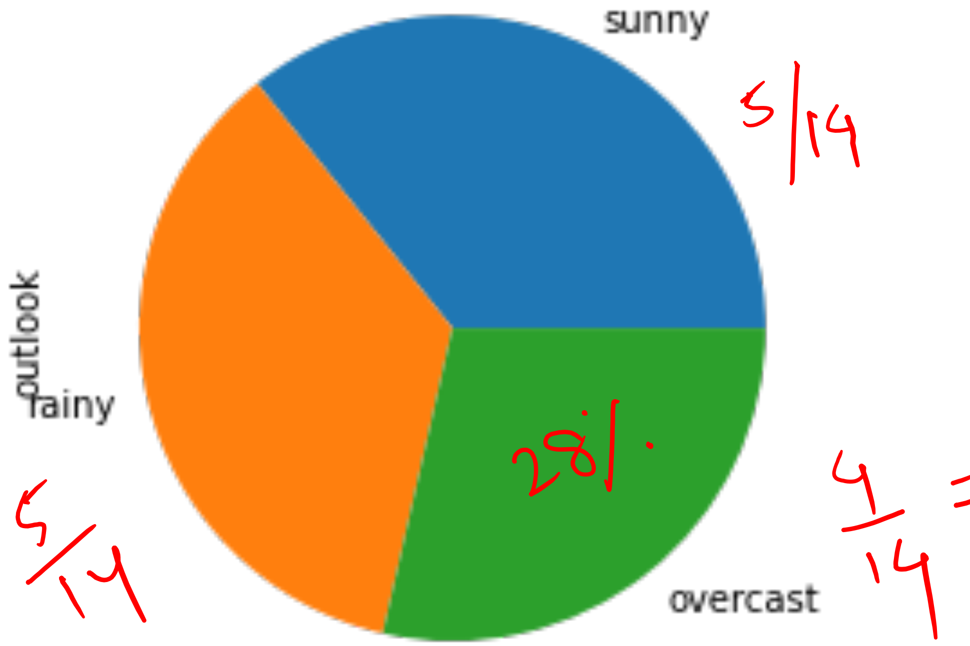
Vertical bar chart

Horizontal bar chart



EXAMPLE: PIE CHART

(percentages)



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.

CONTINGENCY TABLE

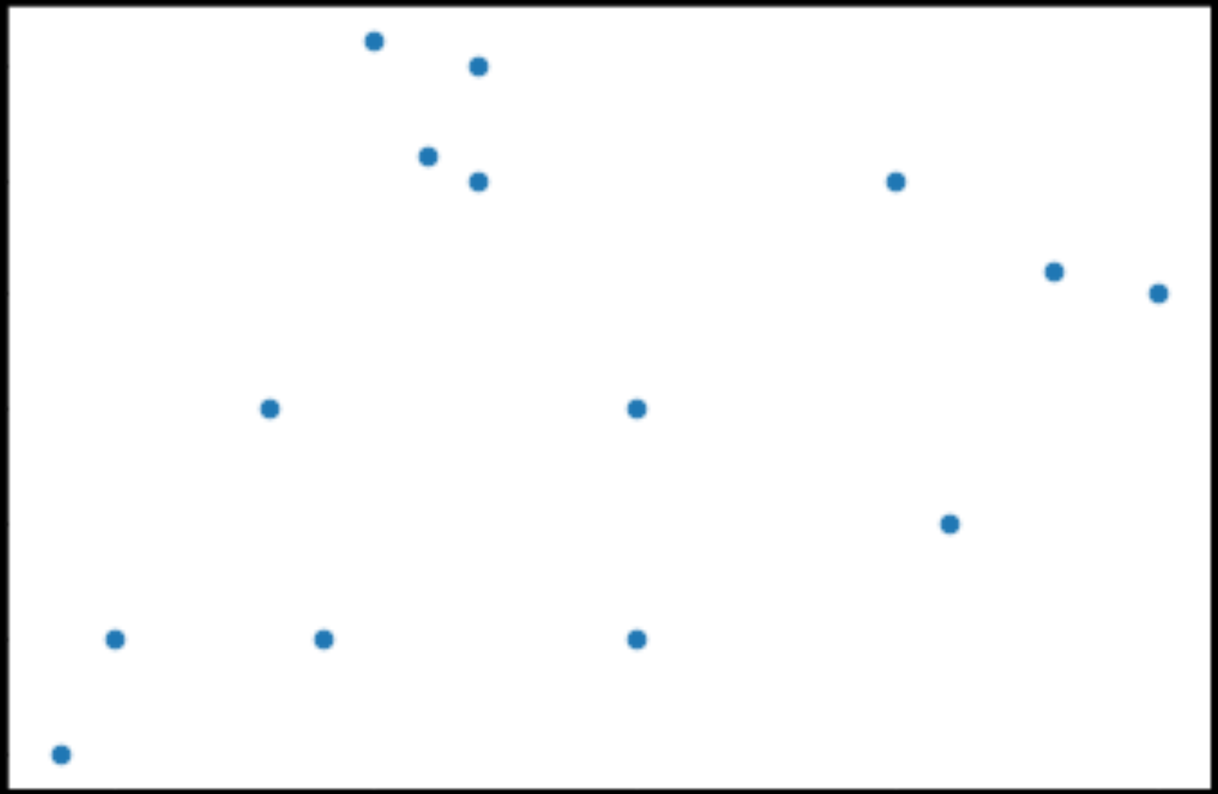
Gender	Designation
Male	Professor
Female	Assoc. Professor
Male	Asst. Professor

Contingency Table:
Gender-wise Breakdown of Different Posts

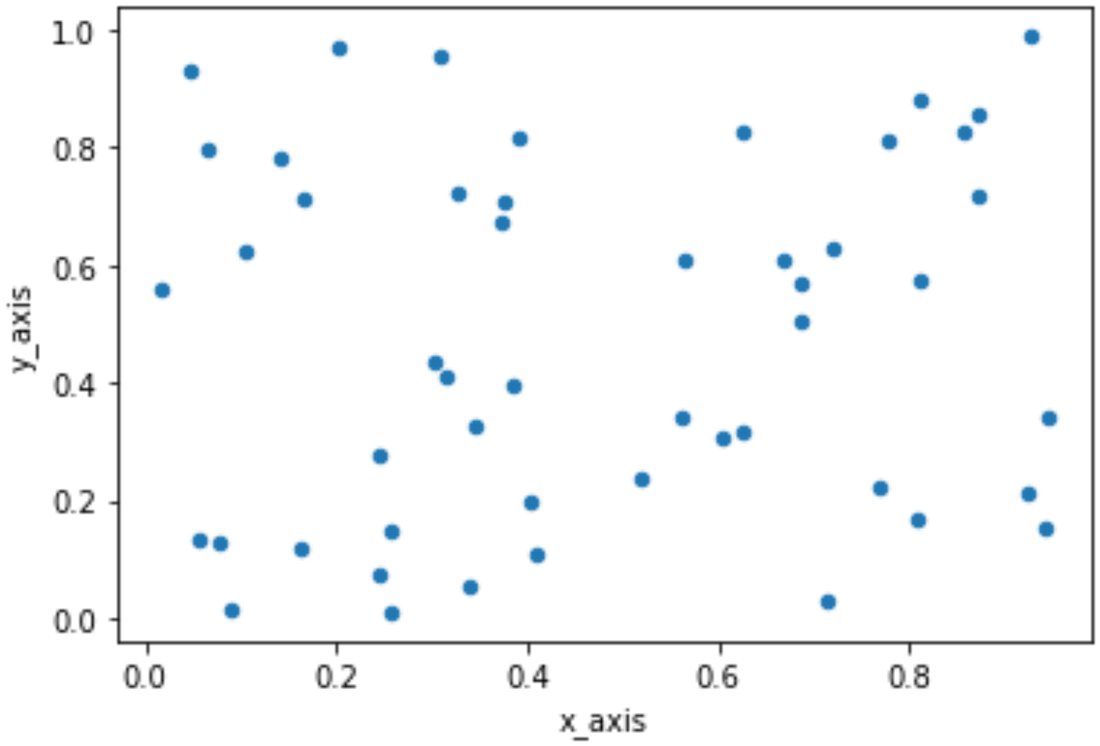
Designation	Gender	
	Male	Female
Professor	3	0
Assoc. Professor	4	1
Asst. Professor	3	0
Sr. Lecturer	4	3
Lecturer	2	2

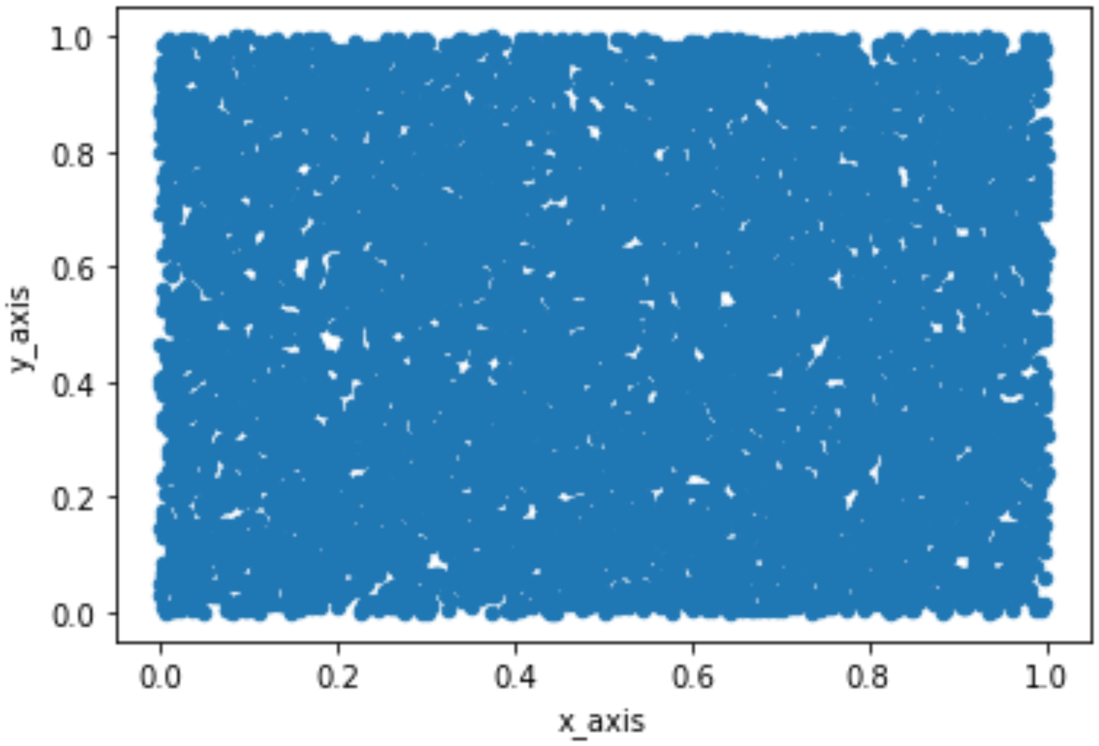
Smoker?	Breathing Problem	
	Yes	No
Non-smoker	1	10
Smoker	4	3

EXAMPLE: SCATTER PLOT



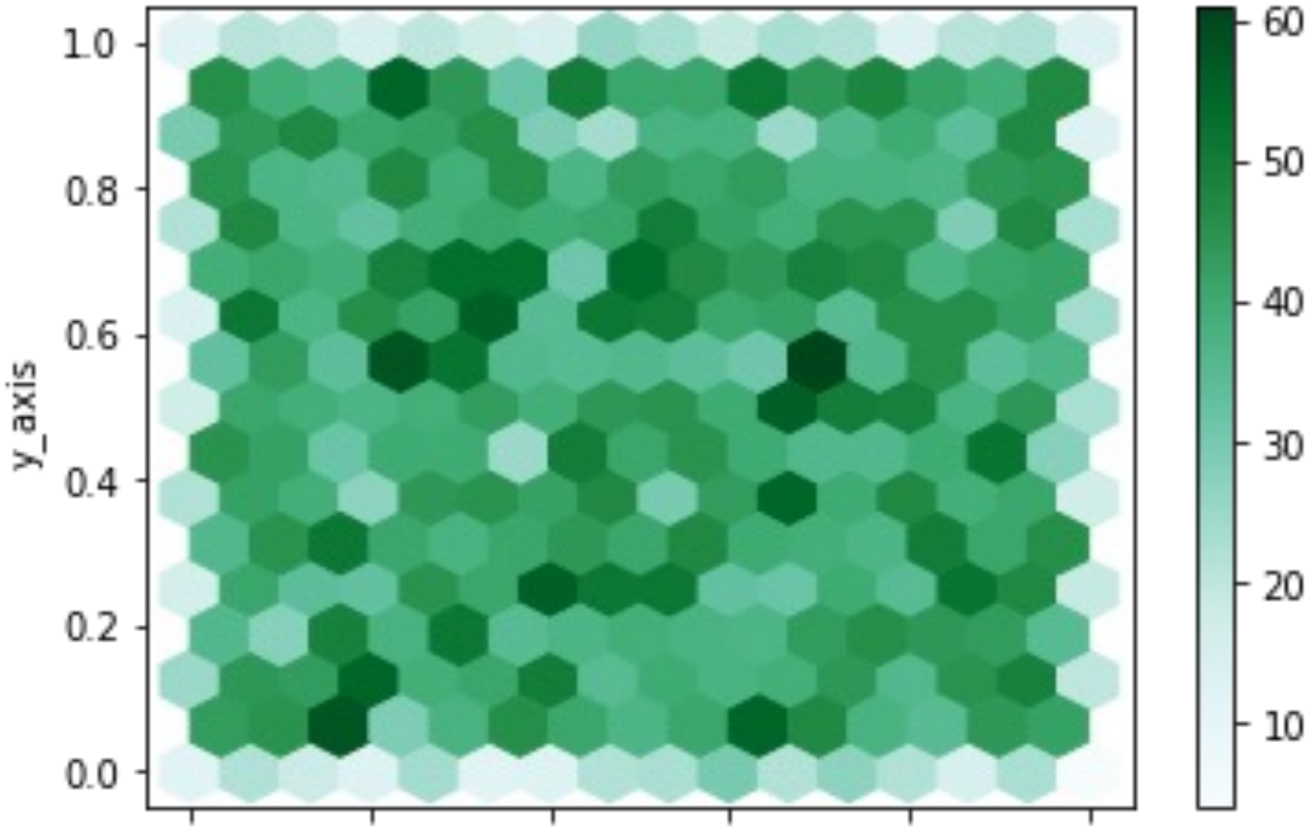
- df1 = pd.DataFrame({'x_axis': np.random.rand(50), 'y_axis': np.random.rand(50)})
- df2 = pd.DataFrame({'x_axis': np.random.rand(10000), 'y_axis': np.random.rand(10000)})





EXAMPLE: HEXAGONAL BINNING

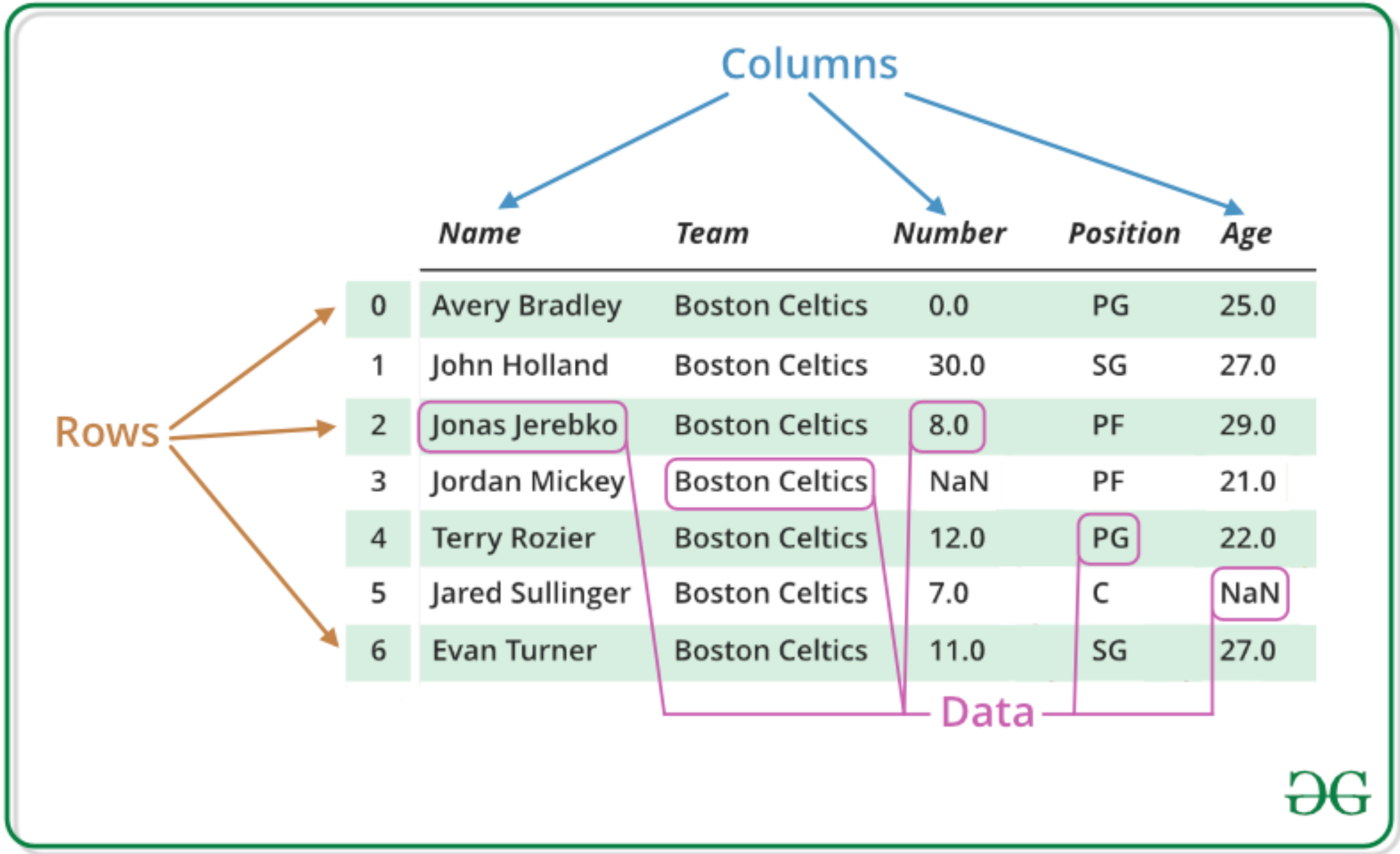
A hexagonal plot is useful for a large dataset. It helps to bin the area of the chart and assigns color intensity based on the frequency on that bin.



Introducing pandas dataframe

- **Pandas DataFrame** is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns).
- A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the **data**, **rows**, and **columns**.

Example: pandas dataframe



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

Creating pandas dataframe using dictionary

```
import pandas as pd
dict1 = {'id':[1,2,3],'name':['alice','bob','charlie'],'age':[20, 25, 32]}
df1 = pd.DataFrame(dict1)
print(df1)
```

Creating pandas dataframe using csv file

```
df = pd.read_csv('../sample_data_1.csv', header = None)
df.columns=['id','state','population','murder_rate’]
print(df)
df.head() # displays first 5 rows
df.tail() # displays last 5 rows
df.count() # displays number of values for each column
```

List of functions on dataframe

Function	Description
count()	number of non-null observations
sum()	sum of values
mean()	mean of values
median()	median of values
mode()	mode of values
std()	standard deviation of values
var()	variance of values
quantile()	quantile of values
min()	minimum value
max()	maximum value
abs()	absolute value
cumsum()	cumulative sum
cumprod()	cumulative product

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