Getting to Know Data & Exploratory Data Analysis

- Imdadullah Khan

Exploratory Data Analysis (EDA): Purpose and Benefits

EDA: Initial investigation of data using summary statistics and graphical representations

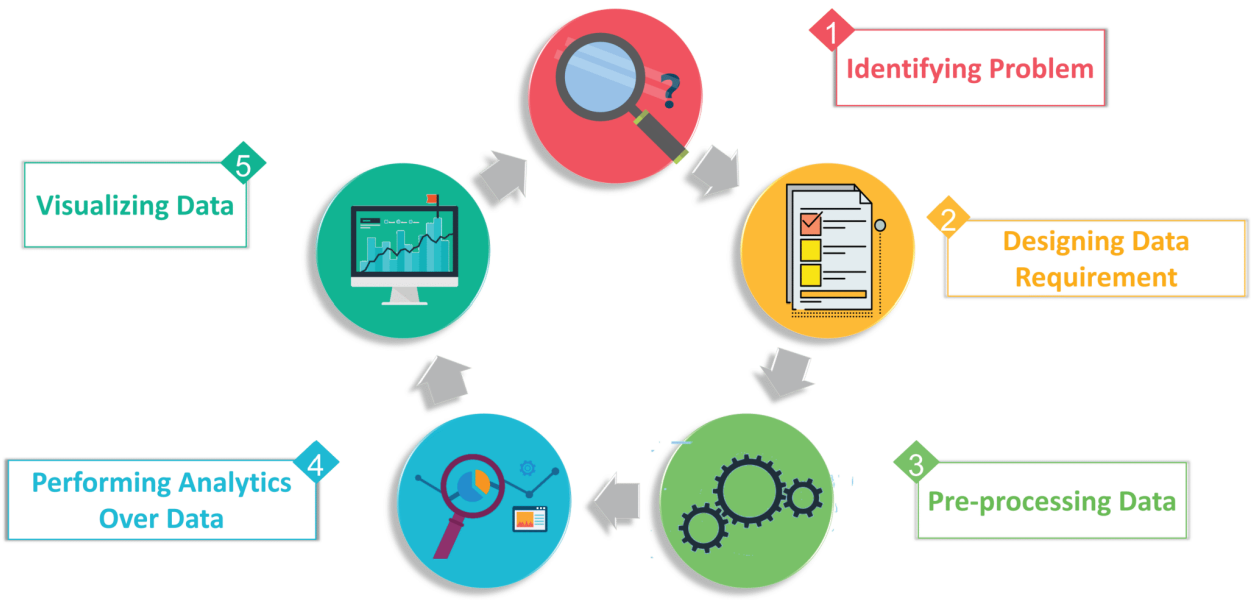
Objectives of EDA are to

- understand data (what it is, where it comes from, what does it represent, kind of values, specific characteristics of data)
- find out if there are missing values? (how to deal with them!)
- spot anomalies (are there outliers?)
- discover patterns (how does the data look like?)
- understand relationships between features (measure similarity, distance and relationship type)
- check our assumptions
- visually describe the data

Exploratory Data Analysis (EDA): Purpose and Benefits

Preliminary exploration and inspection of data is essential for analysis

- It guides preprocessing steps
- Gives a clear picture of data sizes, which helps in selecting the right data structures, tools and even modeling strategies
- Could help reduce data sizes (dimensions or records)



Data object and Attribute

- **Data object**
  - represents an entity in the data set
  - also called data item, data point, instance, example, sample, row, observation
  - e.g. a patient, a movie, a student, a customer, a product, a book, a tweet
  - described by a set of attributes (called feature vector)
- **Attribute**
  - is a data field, representing a feature/characteristic of a data object
  - variable, feature, dimension, column, coordinate, field
  - e.g. reaction to a test, genre/director, course, address, price/category, author, publisher, word

EDA: Size and dimensions of data

- Number and types of attributes
- Dimension of each data item

Sparsity in Data

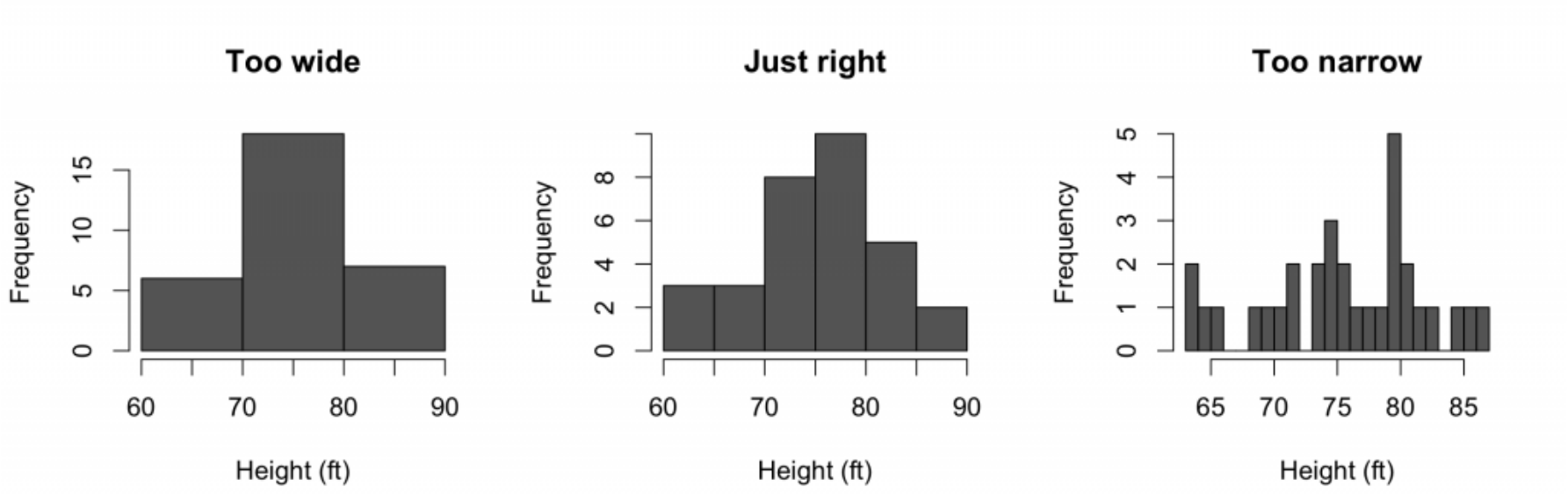
If most of the feature values are missing, then the data is called sparse

- Missing value could be represented as NaN, blank, -, 0
- This could be a problem for many statistical methods
- For efficient computation, can use libraries for sparse data
  - e.g. sparse matrix multiplication, sparse storage schemes

EDA: Resolution of Data

Different resolution reveal different patterns

- If resolution is too fine, a pattern may be buried in noise
- If the resolution is too coarse pattern may disappear
- See number of bins in histograms below



Types of Data

Types of data based on number of attributes

- Univariate Data
- Bivariate Data
- Multivariate Data

EDA: Types of Data

- **Univariate:** Consists of only one feature. Analysis deals with only one quantity that changes.

Heights (cm)
164
167.3
170
174.2
178
180
186

  - What is average height?
  - How much the values deviate form the average height?

EDA: Types of Data

- **Bivariate:** Involves two different features

  Analysis of this type of data deals with comparisons, relationships, causes and explanations

Temperature (°C)	Ice Cream Sales
20	2000
25	2500
35	5000
43	7800

  - Are the temperature and ice cream sales related/dependent?
  - As temperature increases, sales also increases

EDA: Types of Data

- **Multivariate:** Objects are described by more 2 features

  To see if one or more of them are predictive of a certain outcome. The predictive variables are independent variables and the outcome is the dependent variable

Roll #	CS100	SS101	MT200	MGMT240	Major
19100115	A	B	B	C	CS
19100120	B	A	B	C	PHY
19100122	B	B	C	A	CS
19100126	C	A	C	A	EE
19100127	B	A	C	C	CS
19100133	C	B	A	B	PHY
19100135	C	C	A	C	Maths

Types of Attributes

Roll #	Gender	Grade	Age	Major
19100115	Male	B	23	CS
19100120	Male	A	22	PHY
19100122	Female	B	21	CS
19100126	Male	C	19	EE
19100127	Female	A	21	CS
19100133	Female	B	20	PHY
19100135	Male	C	22	Maths

- Nominal/Categorical Attributes
- Ordinal Attributes
- Numeric Attributes

Types of Attributes: Nominal/Categorical

- Possible values are symbols, labels or names of things
  - gender, major, state, color
- Each value represents a category
- They describe a feature **qualitatively** and do not have any order
- **Not quantitative**, arithmetic operations can't be performed on them
  - male - female = ?? green + blue = ??
- Could be coded by numbers (numeric symbols) but they are still symbols, e.g. postal codes, building numbers, roll numbers
- We can compute
  - frequency of each value and the most frequent value
  - ~~middle value~~
  - ~~average value of an attribute~~
- **Special Case Binary:** True/False, Pass/Fail, 0/1
- **Symmetric Binary Attributes:** Both symbols carry the same weight e.g. gender
- **Asymmetric:** Both symbols are not equally important, e.g. Pass/Fail

Types of Attributes: Ordinal Attributes

- Possible values have meaningful order
  - Grades : A,B,C,D
  - Serving Sizes : Small, Medium, Large
  - Ratings : poor, average, excellent
- Can be obtained by discretizing numeric quantities (data reduction)
- No quantified difference between two levels, A is higher/better than B
  - Cannot quantify how much higher is A than B, or
  - if the difference between A and B the same as the difference between B and C
- As the ordering between values exists, we can compute
  - the most frequent value
  - middle value
  - ~~average value of an attribute~~

Types of Attributes: Numeric Attributes

- Quantitative and measurable
- We can quantify the difference between two values
  - temperature, age, number of courses
- **Discrete Numeric Attributes**
  - values of discrete attributes come from a finite or countably infinite sets
  - number of pages, number of students
- **Continuous Numeric Attributes**
  - continuous attributes take real values
  - temperature, height
- We can compute
  - the most frequent value
  - middle value
  - average value of an attribute

- **Interval-Scaled:** No point 0, ratios have no meaning
- **Ratio-Scaled:** A point 0, ratios are meaningful

EDA: Statistical Description of Data

- Estimates that give an overall picture of data
- Summary statistics are numbers that summarize properties of data
- Typical values of variables (features/attributes)
- Spread and distribution of values
- Dependencies and correlations among variables

EDA: Measures of Central Tendencies

- These measures describe the **location** of data
  - location of concentration or middle of data
- Data is distributed around this "center"
- Computed for each attribute
- Three common types of locations
  - Mode
  - Mean
  - Median
- These measures do not give information regarding
  - extreme values in data
  - distribution or spread of the data

EDA: Frequency

Nominal and Ordinal attributes are generally described with frequencies

- The **frequency** of a value is the number of times the value occurs in the dataset
- Some time we use fraction or percentage of time the value appears
  - Probability mass function

Measures of Central Tendencies: Mode

For location of nominal and ordinal attributes one can use the most frequent value

- **Mode** is the most frequent element
- Can have more than one modes
  - unimodal (one mode in data)
  - multi-modal (bimodal, trimodal): more than one modes in data

Not the same as the **Majority** element (a value with frequency more than 50%)

EDA: Measures of Central Tendencies: Mean

For a dataset $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$

(Arithmetic) Mean is the average of the data set

▷ This definition readily extend to higher dimensional data

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Weighted Mean

Weight w_i is associated with value x_i, $\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

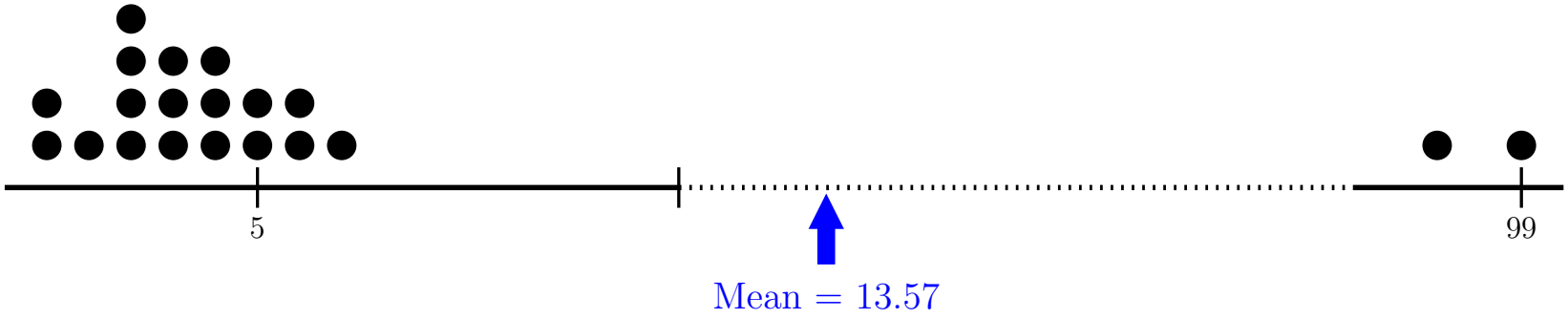
Harmonic Mean $\overline{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

Geometric Mean $\overline{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$

EDA: Other Types of Mean

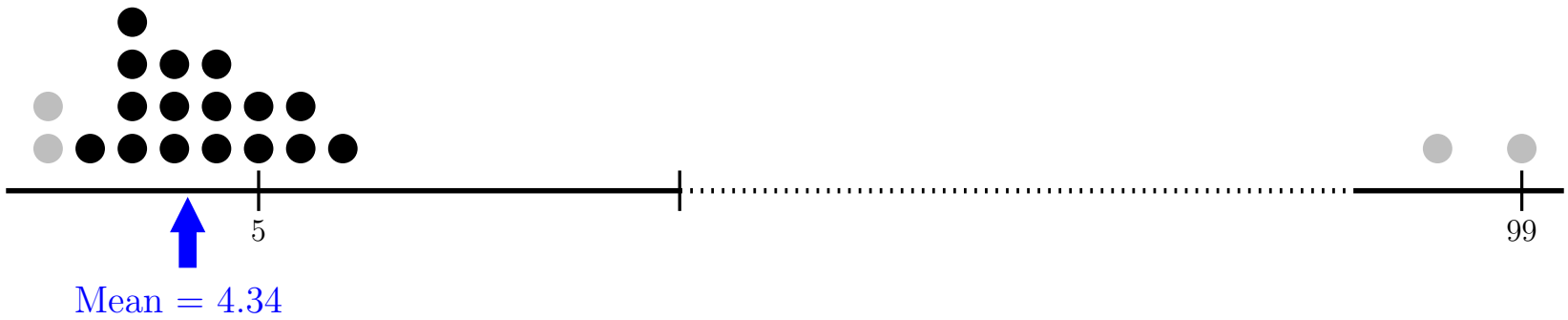
- Arithmetic mean is sensitive to outliers (unstable statistic)
- Just one very high/low value (think $\pm\infty$) makes mean very high/low

2.5 2.5 3 3.5 3.5 3.5 3.5 4 4 4 4.5 4.5 4.5 5 5 5.5 5.5 6 98 99



Trimmed Mean

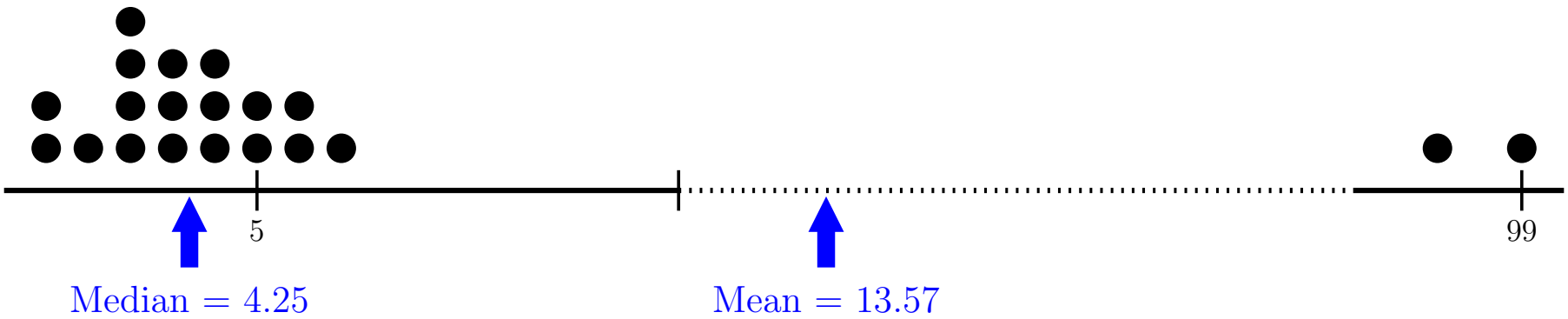
- Ignore $k\%$ of values at both extremes (sorted), then compute mean



EDA: Measures of Central Tendencies: Median

Median is the middle value of a dataset

- Odd/even number of values
- Median is less sensitive to outliers as compared to mean
- Median is good for asymmetric distributions and where data has outliers



- Various possible definitions for median of higher dimensional data
- Mean together variance (see below) has nice properties

EDA: Measures of Spread

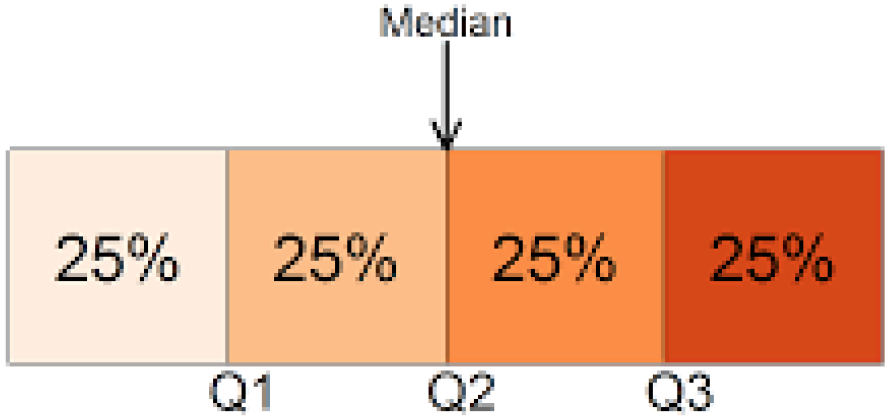
- Location measures do not tell anything about extremes or spread (how extreme are the extremes)
- Measures of spread describe distribution of data

- **Max**
- **Min**
- **Range** := max - min
- **Midrange** := average of min and max
- **Inter-Quartile Range** := 3rd quartile - 1st quartile
- **Mid-spread**
- **Low Spread**
- **High Spread**
- **Variance**
- **Standard Deviation**

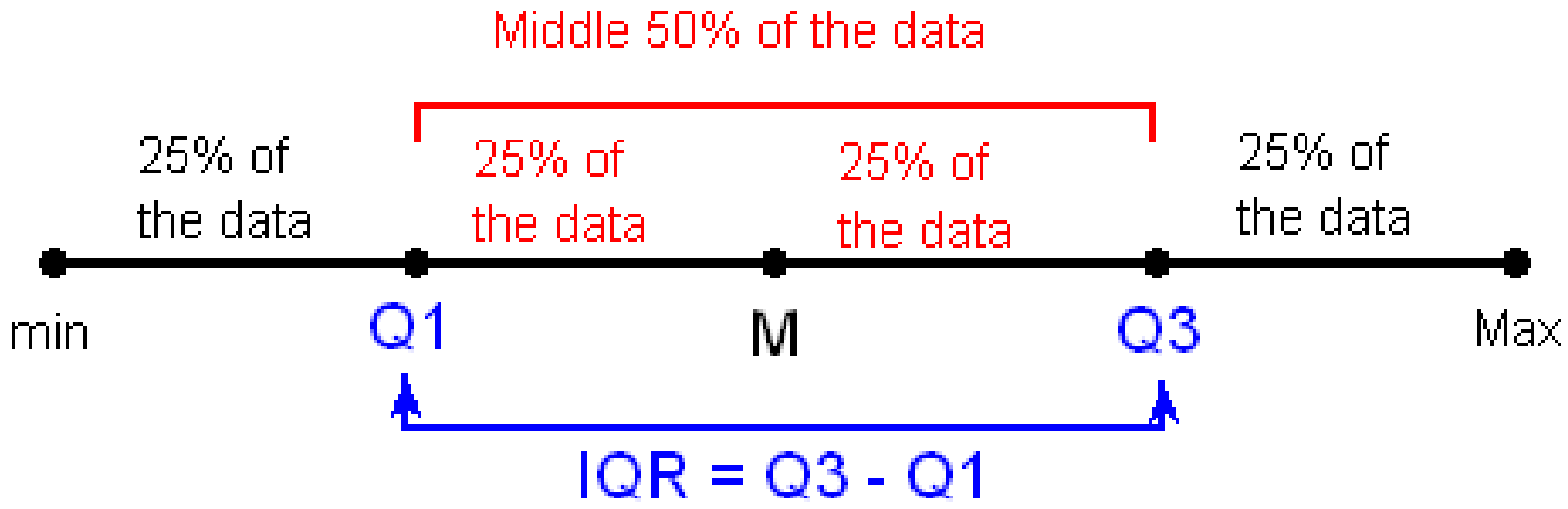
EDA: Quantile

Quantiles are points taken at regular interval so as data is divided into roughly equal sized consecutive subsets

- The ith q-quantile is a data point x such that $\sim {}^{i}/_{q}$ fraction of points are less than x and $\sim {}^{(q-i)}/_{q}$ fraction of points are greater than x
- Median is the first 2-quantile
- 3rd **quartile** := 3rd 4-quantile := 75 **percentile**



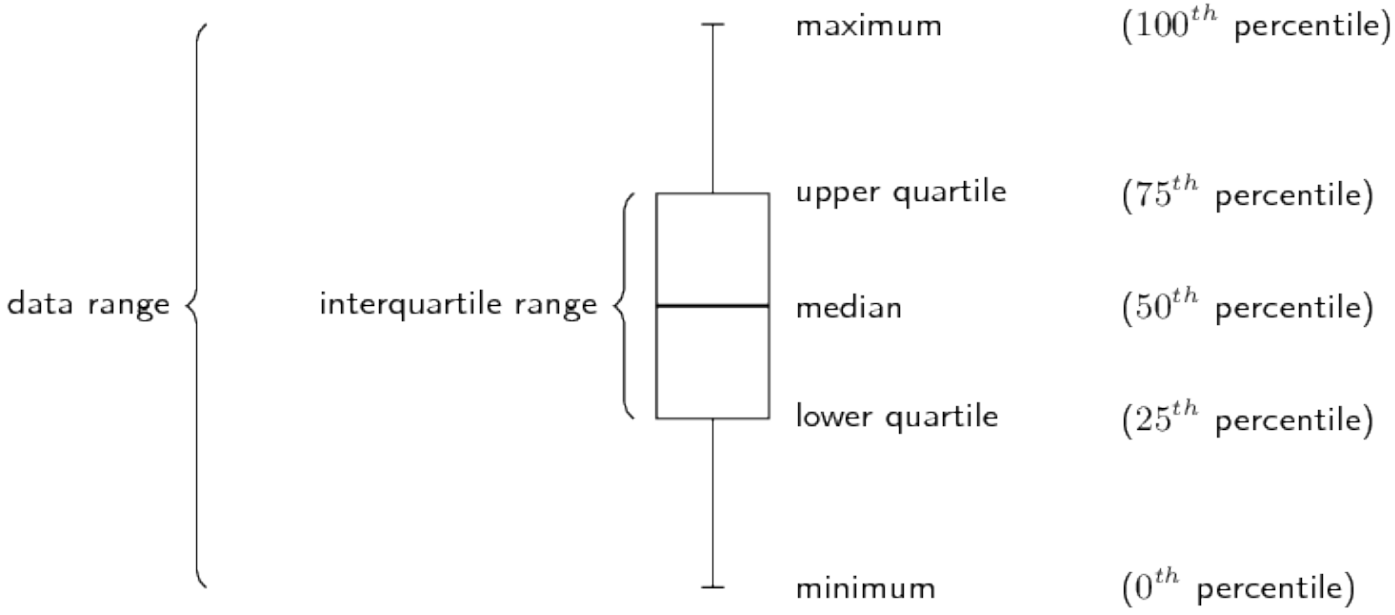
EDA: Measures of Spread



EDA: Five-Number Summary

Five-number summary (elementary EDA of numeric data) consists of

- Min
- 1^{st}/lower quartile
- Median
- 2^{nd}/upper quartile
- Max



EDA: Measures of Spread

Variance

- Measures the deviation in values relative to mean

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

- Squared to avoid cancellation of $+ve$ and $-ve$ deviation
- Average distance from mean could be 0 for data with significant spread
- mean and average distance from mean of both $\{-5, -10, 5, 10\}$ and $\{-100, -50, 50, 100\}$ are 0 and 0
  - there is obviously significantly more spread in the latter data
- Varaince is mean squared deviation from mean
- **Mean Absolute Deviation from mean: MAD** $:= \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n}$
- Variance is easy to compute and has useful mathematical properties

EDA: Measures of Spread

Standard Deviation

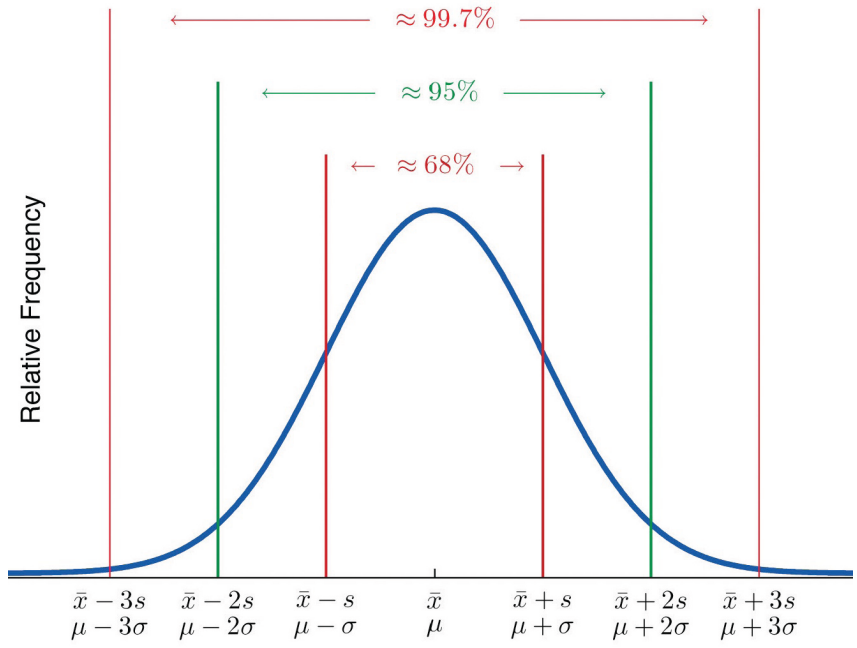
- Variance has different unit than that of original data
- Standard deviation also measures deviation in values relative to mean
- Standard deviation is the square root of variance

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

- Standard deviation restores the measure to the original unit of data

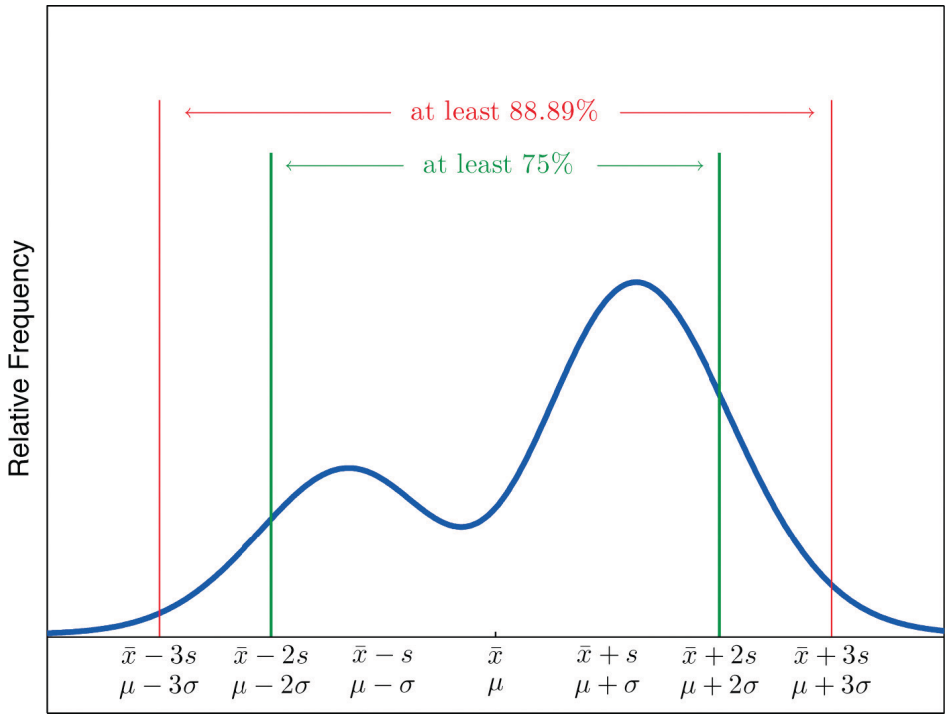
EDA: Normal Distribution (Bell-Curve)

- For normal distribution, there are guarantees that certain number of values must fall within k st-dev from the mean
- At least $\sim 68\%$ must lie within $k = 1$ st-dev $(\overline{x} \pm 1\sigma)$
- At least $\sim 95\%$ must lie within $k = 2$ st-dev $(\overline{x} \pm 2\sigma)$
- At least $\sim 99.7\%$ must lie within $k = 3$ st-dev $(\overline{x} \pm 3\sigma)$



EDA: Three-Sigma Rule - The Empirical Rule

- For any distribution of data, there are guarantees that certain number of values must fall with k st-dev from the mean
- At least $\sim 75\%$ must lie within $k = 2$ st-dev $(\overline{x} \pm 2\sigma)$
- At least $\sim 89\%$ must lie within $k = 3$ st-dev $(\overline{x} \pm 3\sigma)$
- At least $\sim 93\%$ must lie within $k = 4$ st-dev $(\overline{x} \pm 4\sigma)$



EDA: Bivariate Measures

Used for bivariate data or pairs of attributes, more detail later

- **Nominal or Ordinal Attributes**
  - Contingency Table
  - χ^2 statistics
- **Numeric Attributes**
  - Covariance
  - Correlation
  - Correlation Matrix

Contingency Table

- **Contingency table** is used for two nominal or ordinal variables
- Used to determine whether the variable pair is correlated (χ^2-Test)

(nominal) A and B taking values in $\{a_1, a_2, \ldots, a_p\}$ and $\{b_1, b_2, \ldots, b_q\}$

- f_{ij} : frequency of joint occurrence of (a_i, b_j),
  - **observed frequency** of the joint event $(A = a_i, B = b_j)$

Contingency Table:

$C =$

	a_1	a_2	$\ldots$	a_p
b_1				
b_2				
$\vdots$		f_{ij}		
b_q				

	Favor	Neutral	Oppose	f_{row}
Democrat	10	10	30	50
Republican	15	15	10	40
f_{column}	25	25	40	n = 90

χ^2-test for two attributes A and B

χ^2-statistic: A *"correlation"* between two nominal attributes A and B taking values in $\{a_1, a_2, \ldots, a_p\}$ and $\{b_1, b_2, \ldots, b_q\}$

- f_{ij} : frequency of joint occurrence of (a_i, b_j),
  - observed frequency of the joint event $(A = a_i, B = b_j)$
- The expected frequency, e_{ij} of the joint event $(A = a_i, B = b_j)$, under independence assumption
- Estimating probability, $P_{a_i} = Pr\{A = a_i\} = \frac{\sum_{j=1}^{q} f_{ij}}{N}, \quad N = pq$
- $e_{ij} = P_{a_i} \cdot P_{b_j} \cdot N$
- The χ^2 value (Pearson's χ^2-statistics) is $\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
- Large χ^2 values indicates variables are related

Covariance and Correlation

Covariance and correlation are helpful in understanding the dependency/relationship between two numeric variables

- **Covariance** between two variables $x = \{x_1, x_2, \cdots, x_n\}$ and $y = \{y_1, y_2, \cdots, y_n\}$ with means $\overline{x}$ and $\overline{y}$, resp. is defined as

$$COV(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

- Covariance reveals the "proportionality" between variables

- Note when x_i and y_i both are greater or smaller than their respective means, $(x_i - \overline{x})((y_i - \overline{y}))$ is positive and vice-versa
- $COV(x, y) < 0 \implies$ inverse proportionality
- $COV(x, y) > 0 \implies$ direct proportionality
- $COV(x, y) = 0 \implies$ no linear relation

Covariance and Correlation

Some properties of covariance that readily follow from definition

- $COV(x, y) = COV(y, x)$
- $COV(x, x) = Var(y, x)$
- If x and y are independent, then $COV(x, y) = 0$
- For constant a and b
  - $COV(x, a) = 0$
  - $COV(ax, by) = ab\ COV(x, y)$
  - $COV(x + a, y + b) = COV(x, y)$
- $COV(x, y + z) = COV(x, y) + COV(x, z)$

Covariance and Correlation

Correlation

- covariance value depends on magnitude and scale of variable x and y
- Correlation quantifies how strongly two variables are linearly related

$$r_{xy} = corr(x, y) = \frac{COV(x, y)}{\sigma_x . \sigma_y}$$

- $corr(x, y) \in [-1, 1]$
- It is not affected by changes in scale of variables x and y
  - $corr(x, y) = -1 \implies$ perfect negative linear association
  - $corr(x, y) = 1 \implies$ perfect positive linear association
  - $corr(x, y) = 0 \implies$ no linear association

Correlation

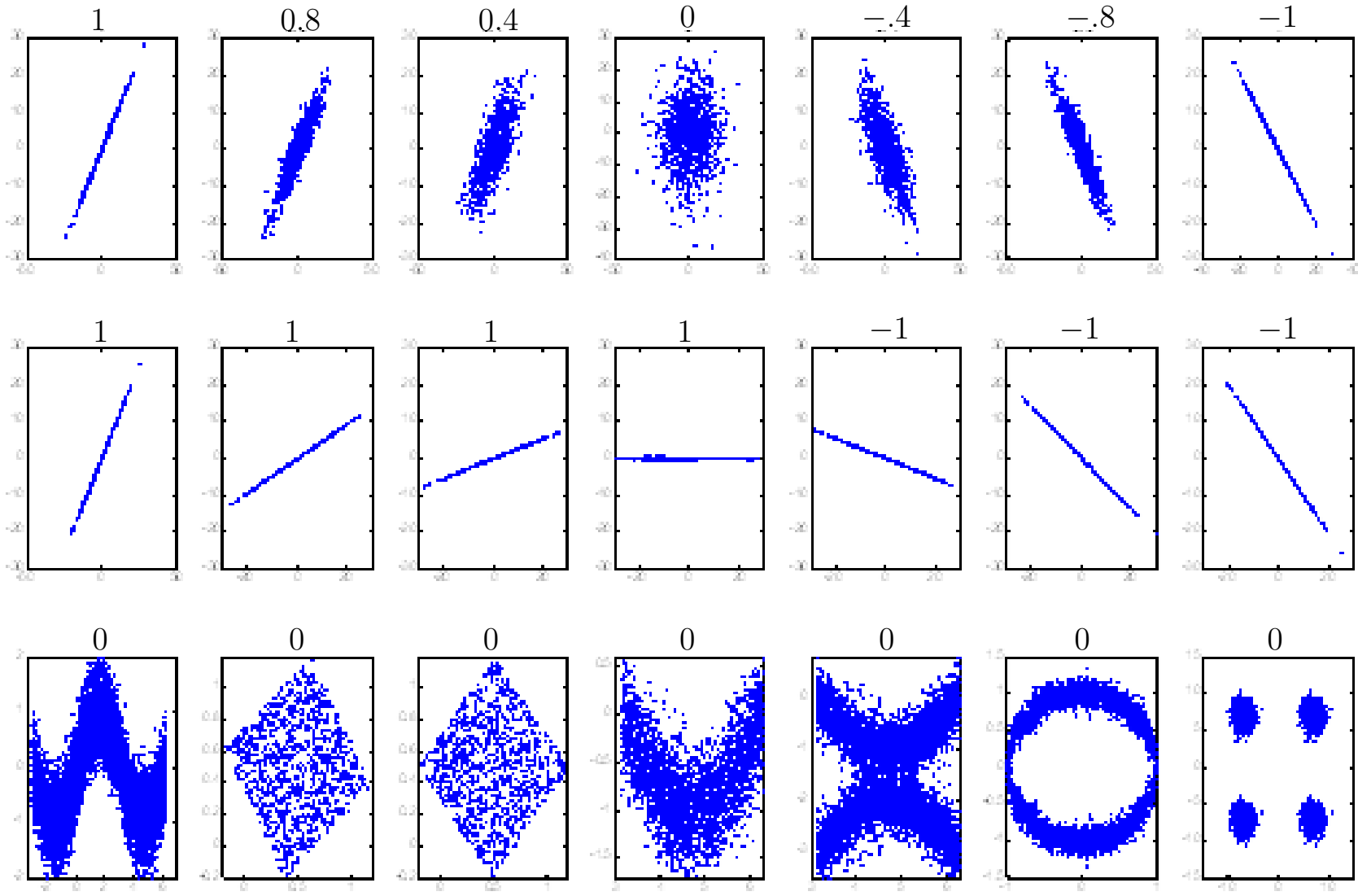


Figure: x and y-axis represent variables - their correlations is on the top

Correlation matrix

- A table of pairwise correlation coefficients between variables
- Each cell shows the correlation between two variables
- Used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses

	body weight in kg	total sleep (hours/day	sleep exposure index (1-5)	maximum life span (years)
body weight in kg	1	-.307	.338	.302
total sleep (hours/day)	-.307	1	-.642	-.410
sleep exposure index (1-5)	.338	-.642	1	.360
maximum life span (years)	.302	-.410	.360	1

Diagrammatic Representations of Data

- Easy to understand: Numbers do not tell all the story. Diagrammatic representation of data makes it easier to understand
- Simplified Presentation: Large volumes of complex data can be represented in a simplified and intelligible diagram
- Reveals hidden facts: Diagrams help in bringing out the facts and relationships between data not noticeable in raw/tabular form
- Easy to compare: Diagrams make it easier to compare data

Types of Diagrams

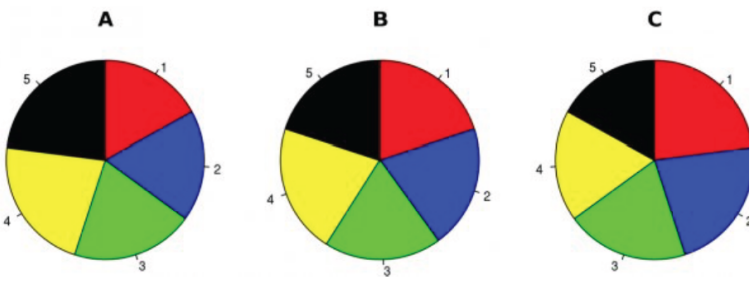
We will briefly discuss and use the following types of diagrams

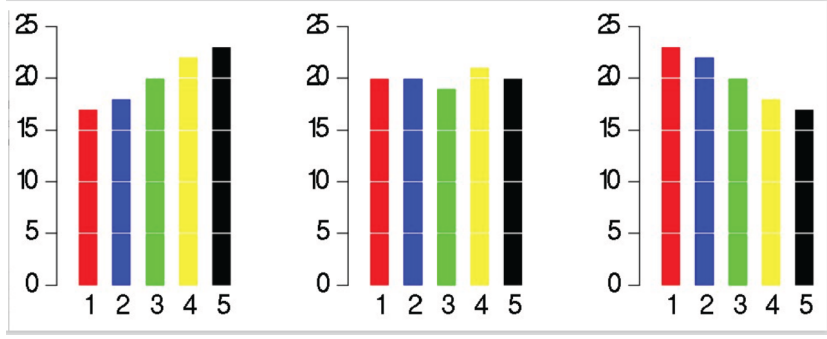
▷ More on importance of visualization later

- Bar Charts
- Histogram ▷ and also overlapping histogram
- Box Plot ▷ and also side-by-side box-plots
- Scatter Plot ▷ and scatter plot matrix
- Heat map
- Line Graph
- Parallel Axis Plot
- Word-Cloud

Bar charts

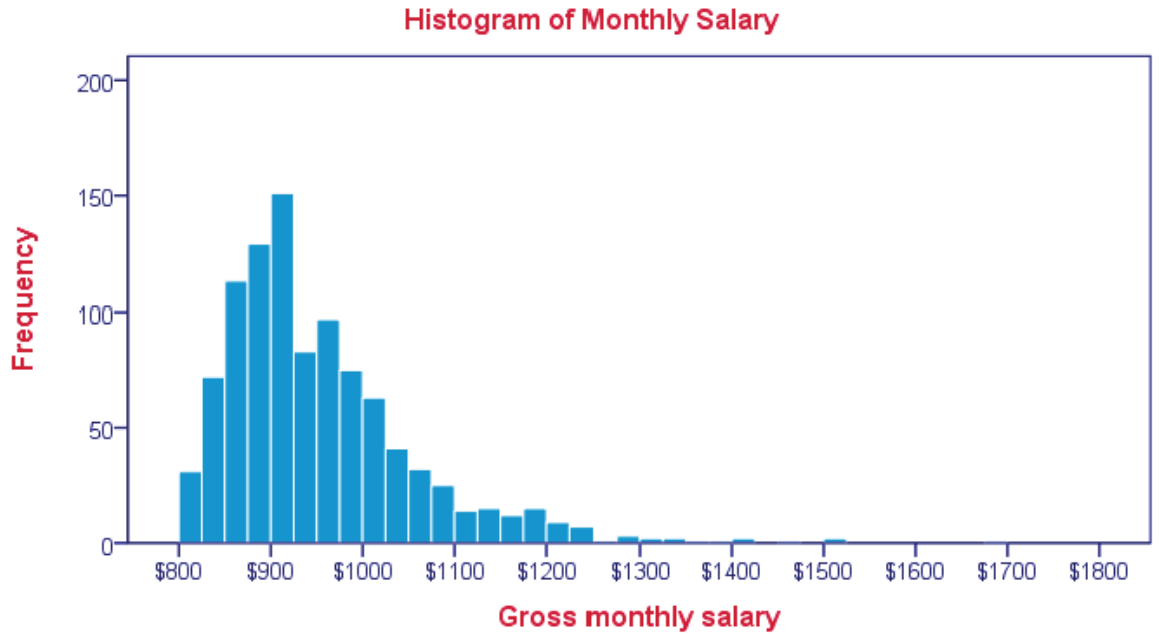
- Generally used for a nominal and ordinal variables
- Different bars (usually colored/shaded differently) for distinct values (levels, categories, symbols) of the variable
- Height of bar represent frequencies of each symbol (value)
- Can reveal variables that have no or limited information e.g. constants
- Note that we can use pie charts for the same purpose too
- Humans perceive difference in lengths better than in angles





Histograms

- Represent distribution of data in a numeric/continuous variable (estimates probability distribution of a numeric variable)
- Group values by a series of intervals (bins - usually consecutive non-overlapping subintervals covering range of data)
- Plot the number of values falling in each bin (represented by the height of the bar)
- Normalized histogram shows proportion of values in each bin



Histograms

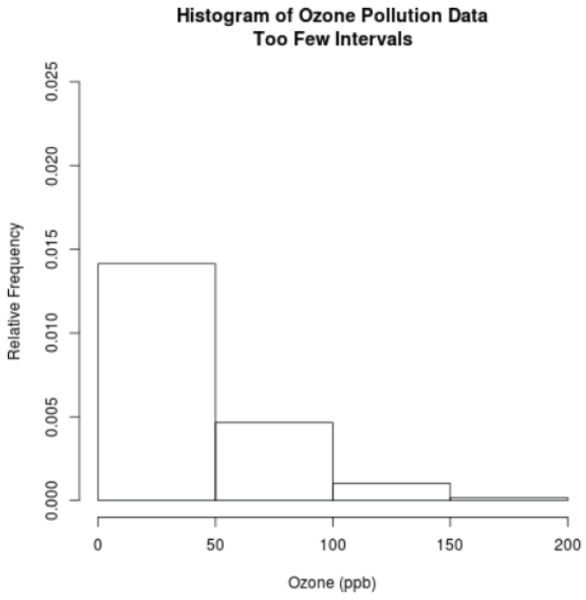
A histogram with appropriate number/length of bins reveals

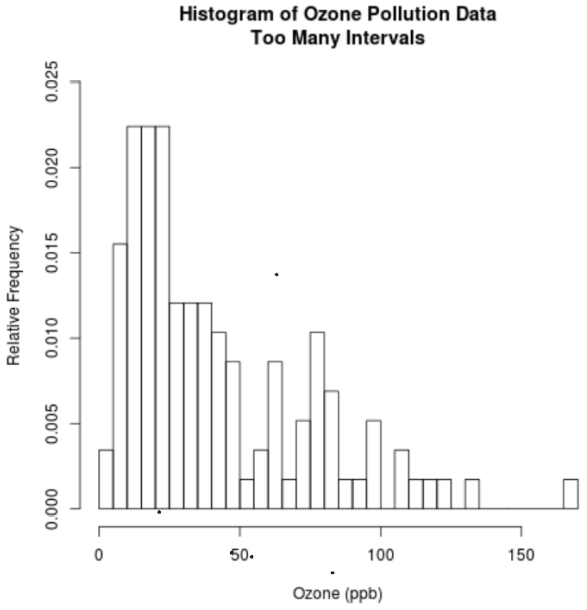
- Where are the data located
- Where/What are the extremes
- What is the distribution of the data
- How the data is spread out
- If the distribution is symmetric, or have skew (left or right)

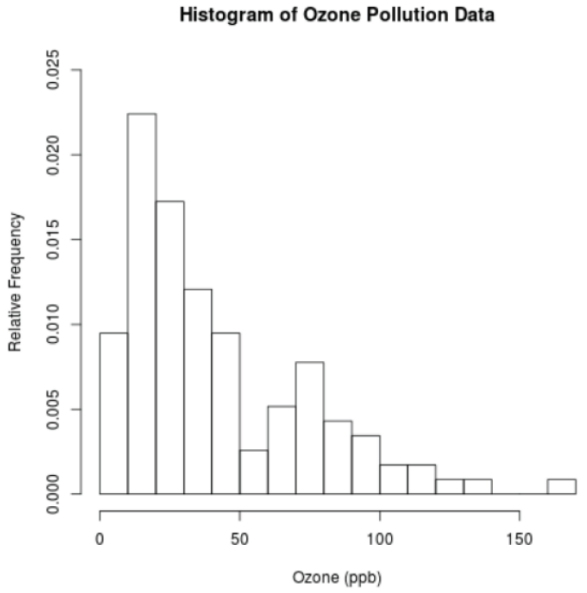
Histograms

- Number and sizes of bins are important considerations
- Bins do not have to be of equal sizes
- For unequal bin sizes height of the bar is not the frequency of values in the bin, it is the **frequency density**
  - Area of the bar is proportional to the frequency
  - Number of items per unit of the variable of x-axis
- Too many bins in histogram gives too much unnecessary details (shows too much noise)
- Too few bins give almost nothing, obscure the underlying patterns

Histograms

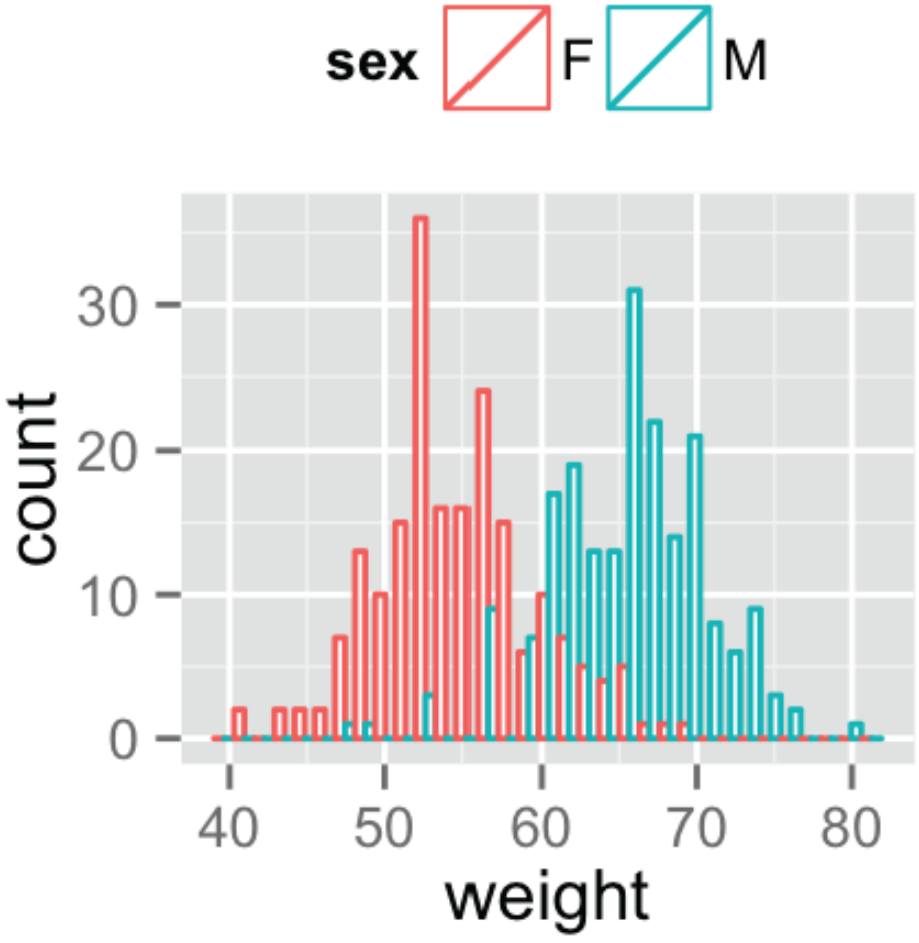

Histogram of Ozone Pollution Data
Too Few Intervals
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200
Ozone (ppb)


Histogram of Ozone Pollution Data
Too Many Intervals
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
Ozone (ppb)


Histogram of Ozone Pollution Data
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
Ozone (ppb)

Overlapping Histograms

Useful in observing distribution of values with respect to a nominal variable



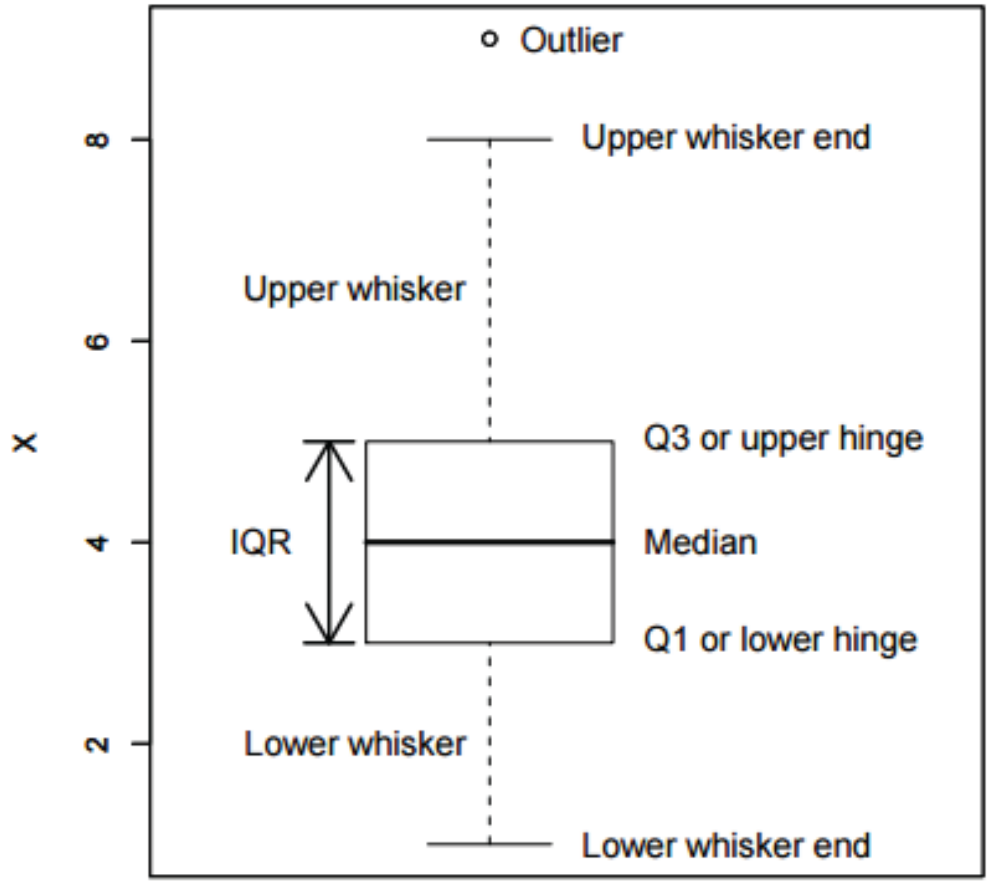
Box Plots

Another way of displaying the distribution of data (somewhat)

- Top and bottom lines of the box are 3rd and 1st quartiles of data
- Length of the box is the inter-quartile range (midspread)
- The line in the middle of the box is median of data
- The top whisker denotes the largest value in the data that is within 1.5 times midspread ($Q3 \times 1.5 \cdot IQR$)
- Similarly the bottom whisker
- Anything above and below the whiskers are considered outliers
- Relative location of median within the box tells us about data distribution
- We find out at what end are the outliers if any

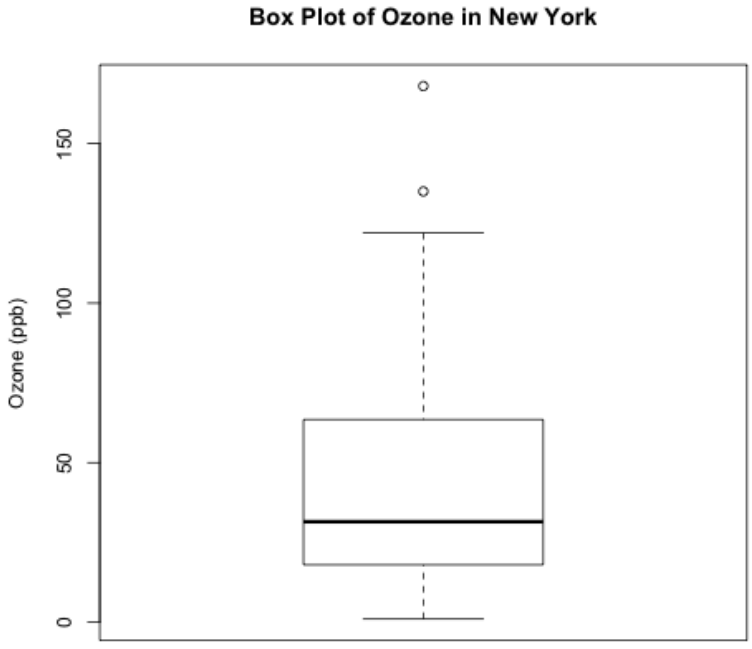
Box Plots

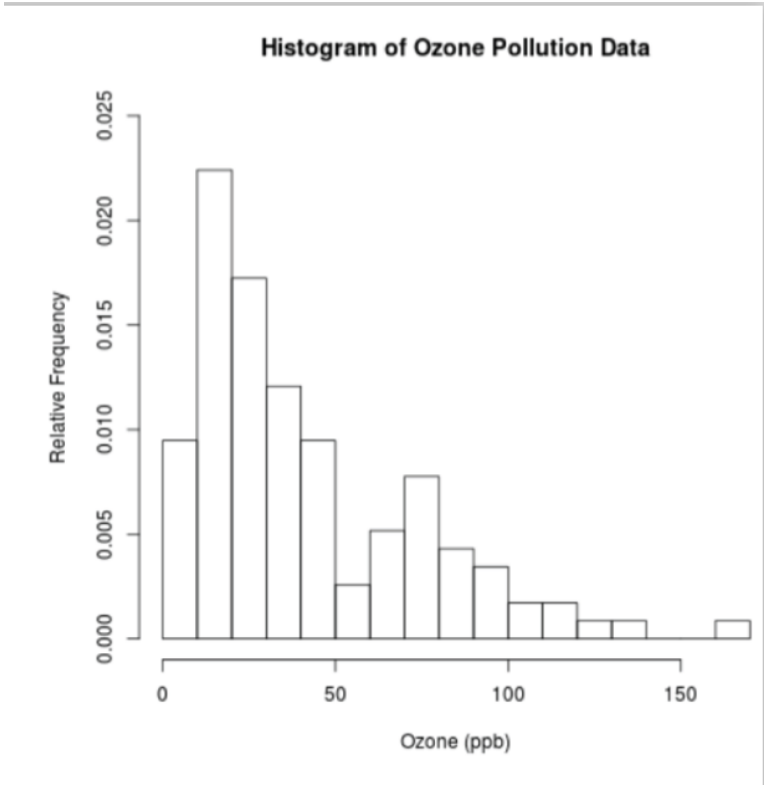
Box-Plots or Box and Whisker diagrams



Box Plots

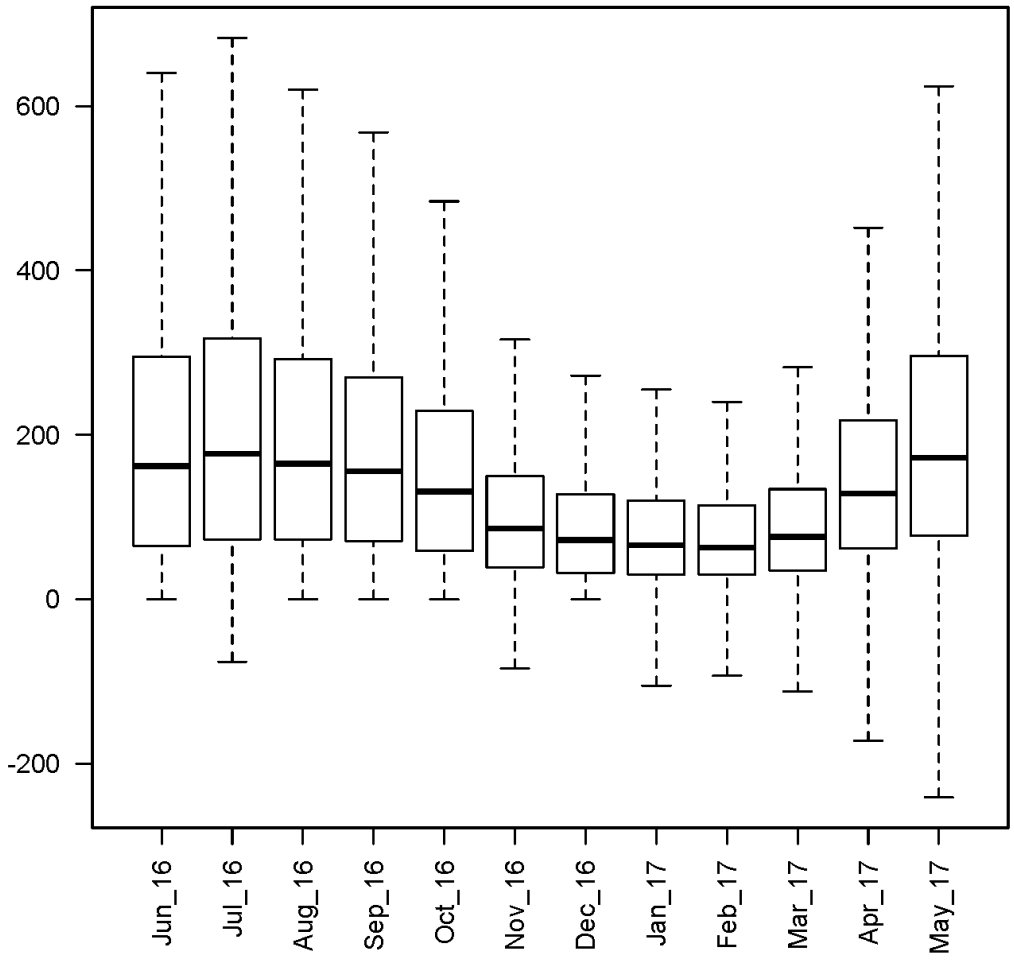
- Can get some idea of skew by observing the shorter whisker
- Various norms for whiskers (sometime) top whisker is 90th percentile
- Uni-modality and multi-modality type information is generally not clear from box plots





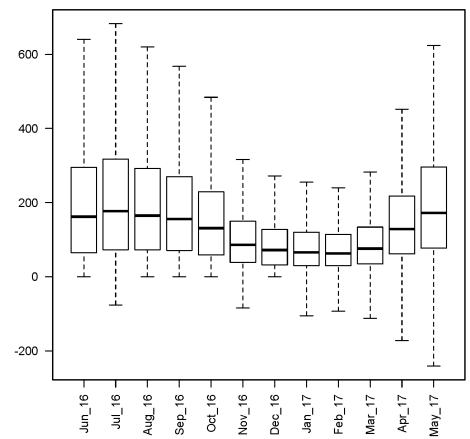
Side-by-side Box Plots

- Extremely useful for comparisons of two or more variables.
- To compare numeric variables, we draw their box-plots in parallel



Side-by-side Box Plots

- Side by side groupwise box plots are extremely useful
- Groups are based on values of a categorical variable
- It reveals whether a factor (the categorical variable) is important
- It addresses whether the location of data differ between groups
- To some extent it also reveals whether distribution and variation differ between groups
- Overlapping histograms are more suitable for the latter question, unless there is too much overlap

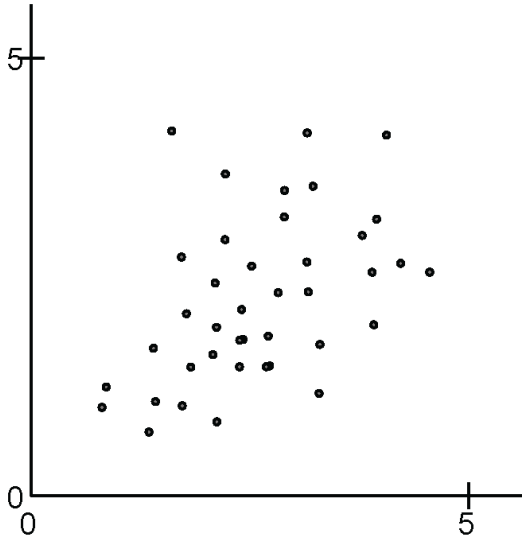


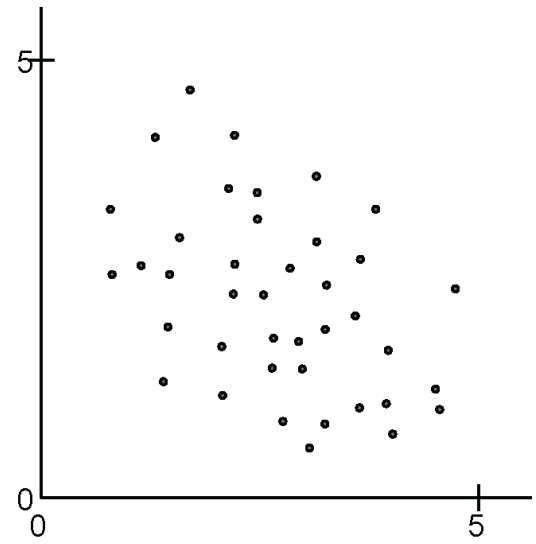
Bivariate data: Scatter Plot

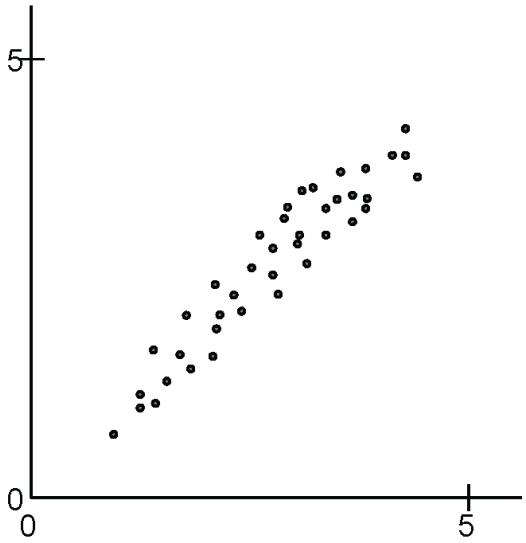
Scatter Plot is the best visual

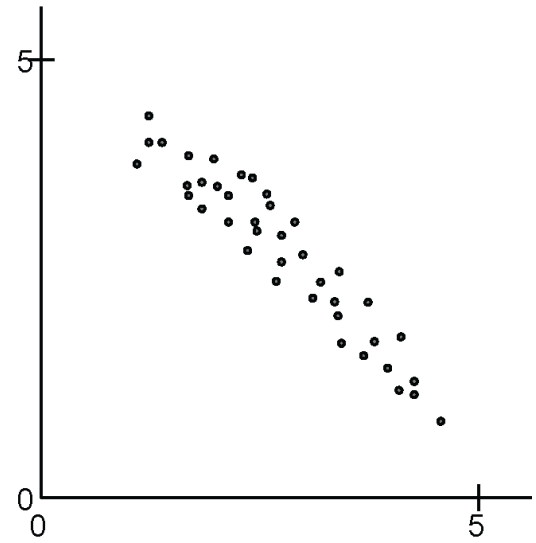
- When we want to visualize two dimensional data
- or want to see how two variables are related to each other
- This directly represent the two dimensional observations as points in $\mathbb{R}^2$. Plot one variable on x-axis and other on y-axis
- It reveals correlations among the variables
- If one or both variables are highly skewed, then scatter plots are hard to examine, as bulk of the data is concentrated in a small part of plot
- For this we should use some kind of transformation, explained later on one or both the variables
- log-scaled plots can also be used in such cases

Bivariate data: Scatter Plot



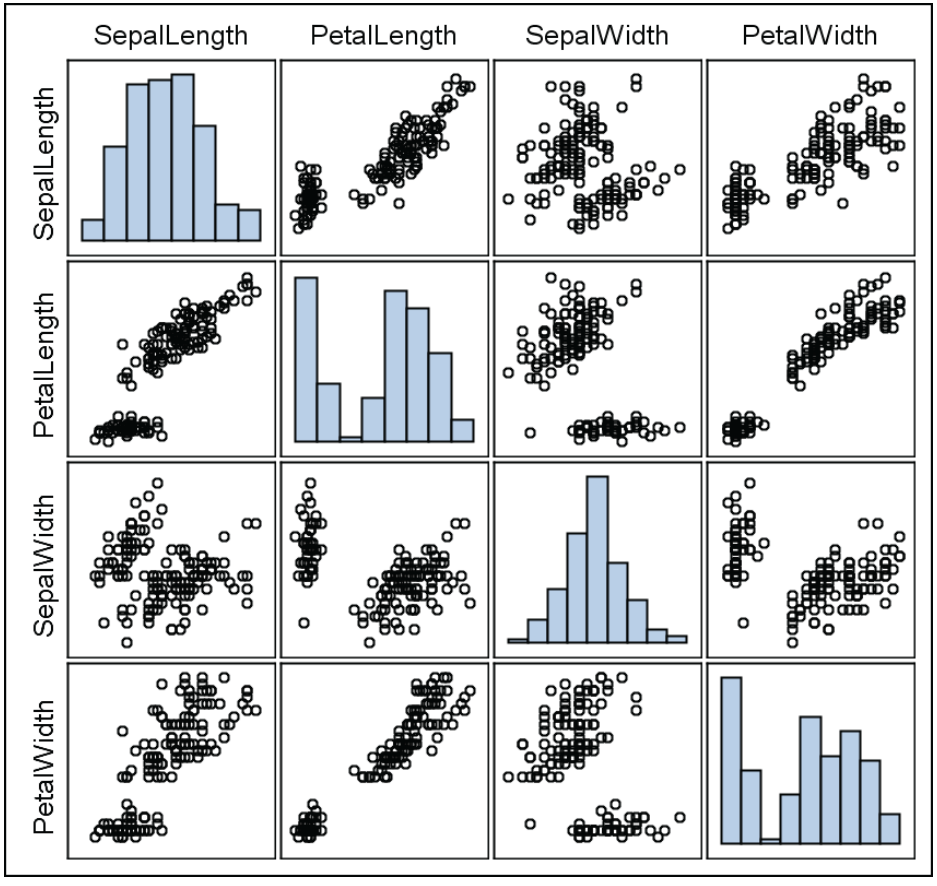






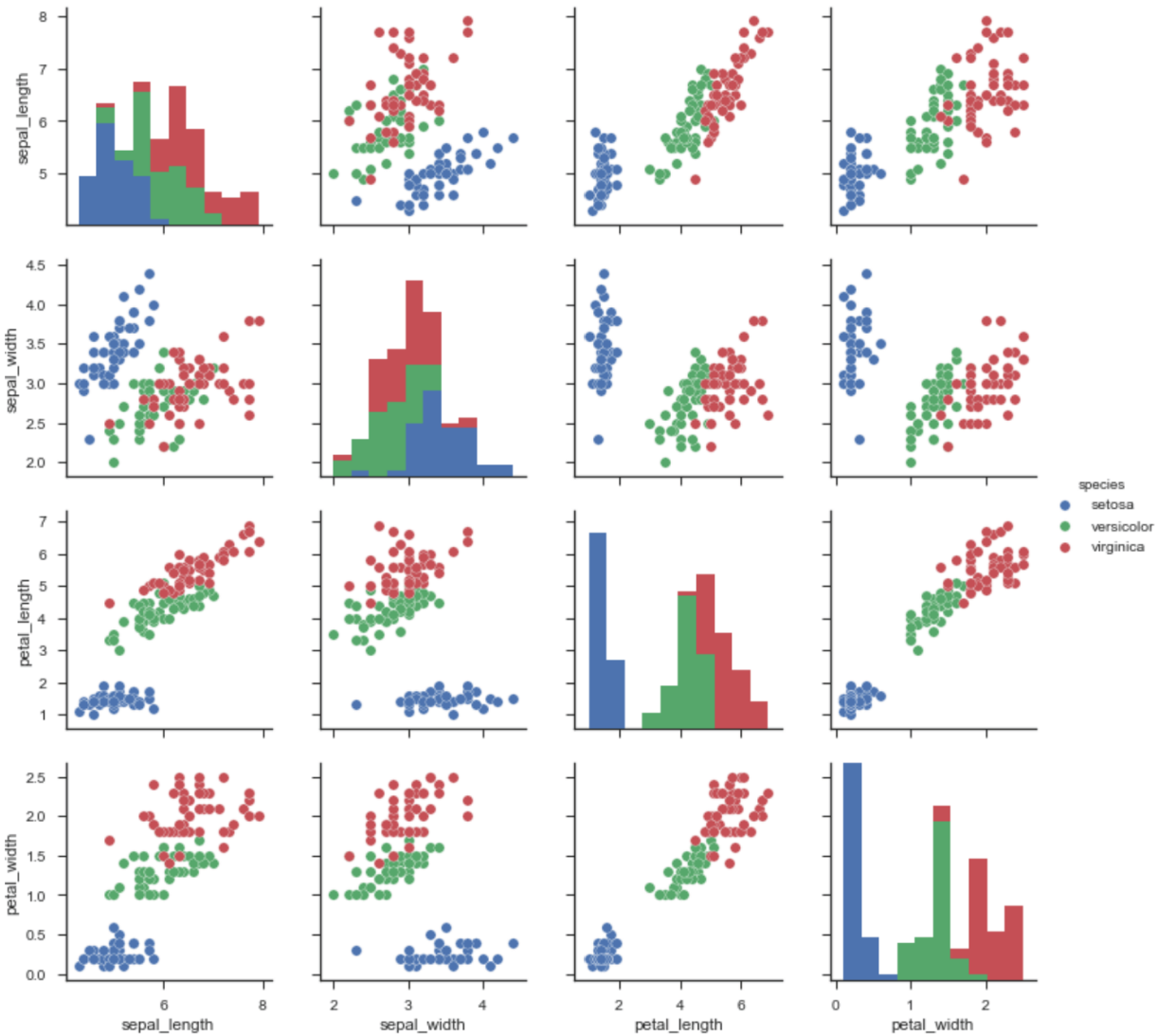
Scatter Plot Matrix

- Pairwise scatter plots, pairwise correlations and individual histograms or density plots
- Summarize the relationships of all pairs of numerical attributes



Scatter Plot Matrix

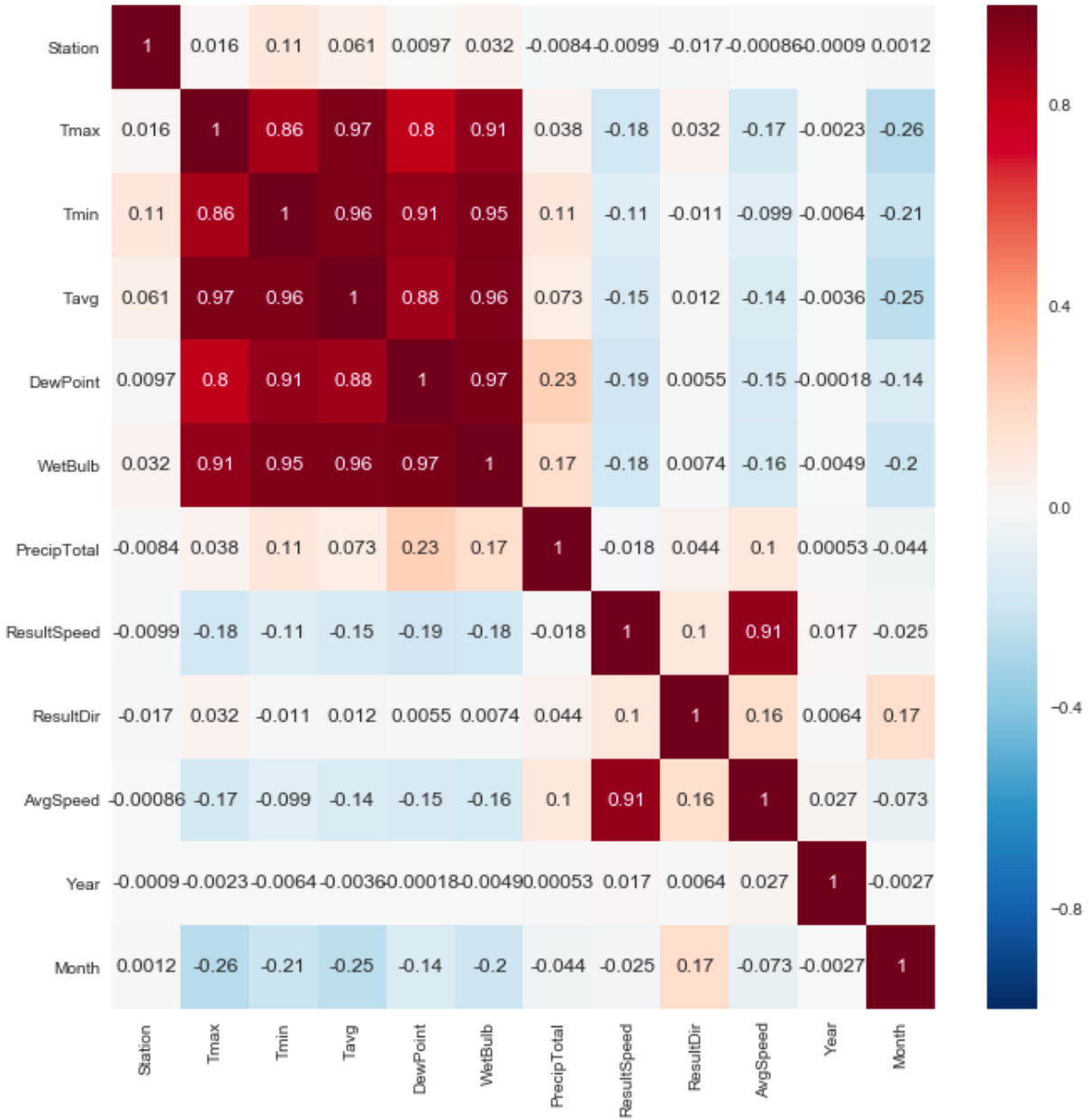
- Scatter plot (matrix) can be combined with information in a nominal attribute encoded through color or marker shape



Multivariate data: Heat Map

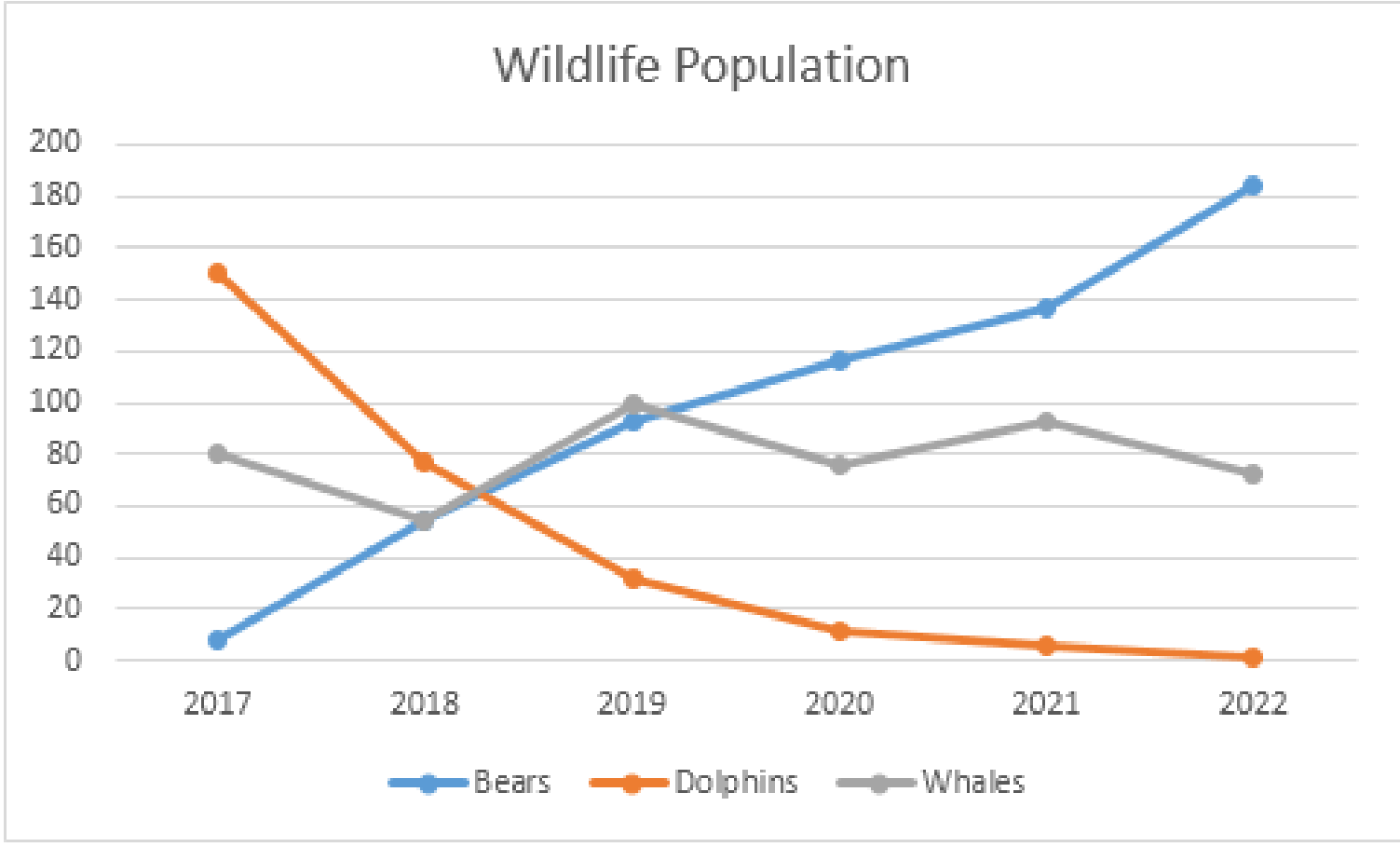
- This is plot to present relationship between multivariate data
- Provides a numerical value of the correlation between each variable
- Also provides an easy to understand visual representation of those numbers (colors shades)
- Darker red showing high correlation
- Dark blue showing none or negative correlation
- Can be used to visualize any matrix

Heat Map

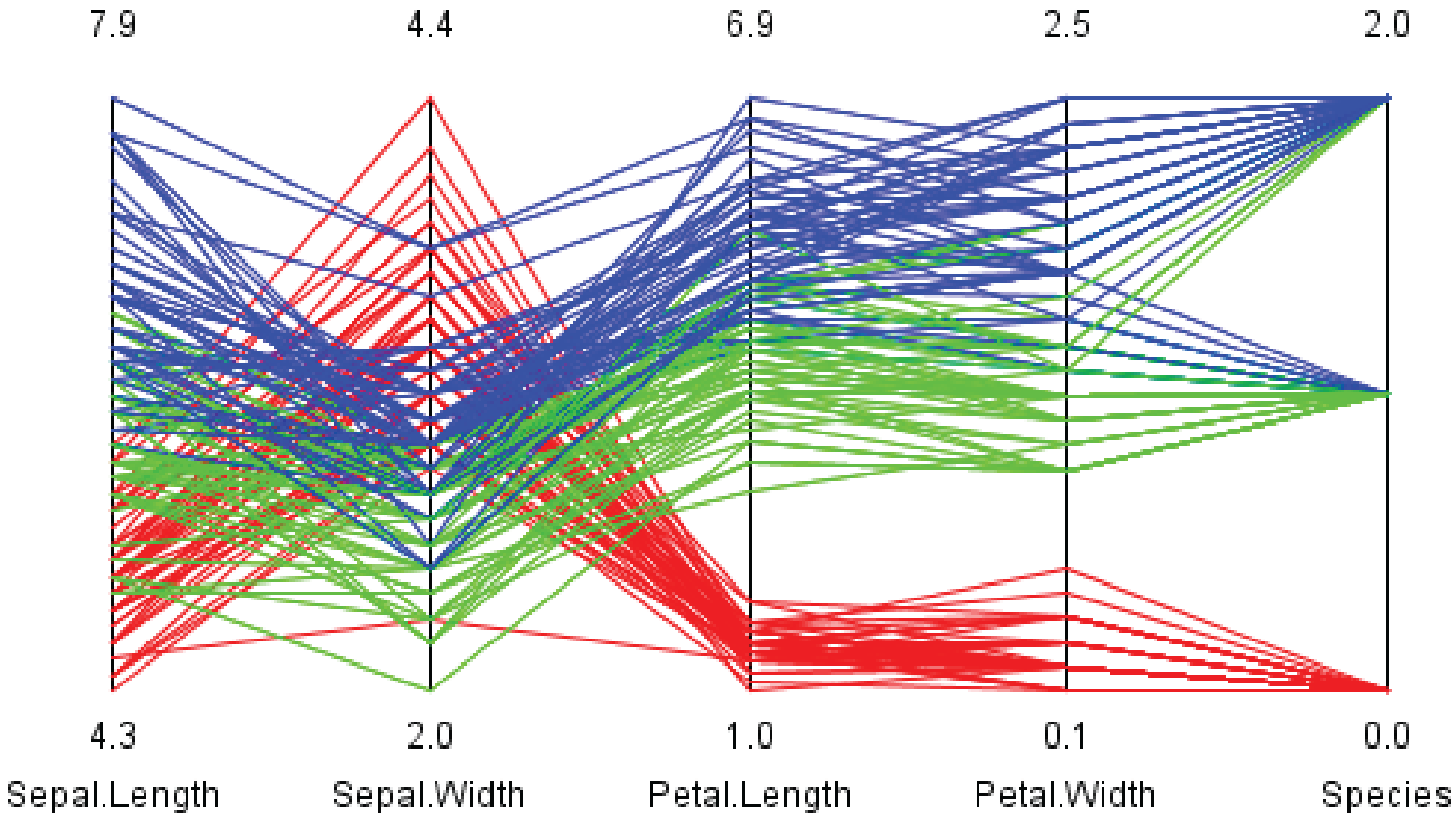

Station 1 0.016 0.11 0.061 0.0097 0.032 -0.0084 -0.0099 -0.017 -0.00086 -0.0009 0.0012
Tmax 0.016 1 0.86 0.97 0.8 0.91 0.038 -0.18 0.032 -0.17 -0.0023 -0.26
Tmin 0.11 0.86 1 0.96 0.91 0.95 0.11 -0.11 -0.011 -0.099 -0.0064 -0.21
Tavg 0.061 0.97 0.96 1 0.88 0.96 0.073 -0.15 0.012 -0.14 -0.0036 -0.25
DewPoint 0.0097 0.8 0.91 0.88 1 0.97 0.23 -0.19 0.0055 -0.15 -0.00018 -0.14
WetBulb 0.032 0.91 0.95 0.96 0.97 1 0.17 -0.18 0.0074 -0.16 -0.0049 -0.2
PrecipTotal -0.0084 0.038 0.11 0.073 0.23 0.17 1 -0.018 0.044 0.1 0.00053 -0.044
ResultSpeed -0.0099 -0.18 -0.11 -0.15 -0.19 -0.18 -0.018 1 0.1 0.91 0.017 -0.025
ResultDir -0.017 0.032 -0.011 0.012 0.0055 0.0074 0.044 0.1 1 0.16 0.0064 0.17
AvgSpeed -0.00086 -0.17 -0.099 -0.14 -0.15 -0.16 0.1 0.91 0.16 1 0.027 -0.073
Year -0.0009 -0.0023 -0.0064 -0.0036 0.00018 -0.0049 0.00053 0.017 0.0064 0.027 1 -0.0027
Month 0.0012 -0.26 -0.21 -0.25 -0.14 -0.2 -0.044 -0.025 0.17 -0.073 -0.0027 1
Station Tmax Tmin Tavg DewPoint WetBulb PrecipTotal ResultSpeed ResultDir AvgSpeed Year Month
0.8
0.4
0.0
−0.4
−0.8

Line graphs

- Line graphs are used for time series e.g. player's yearly average, student's semester gpa or hourly energy consumption
- Two or more time series can be compared in different colors or markers (legend should be provided)



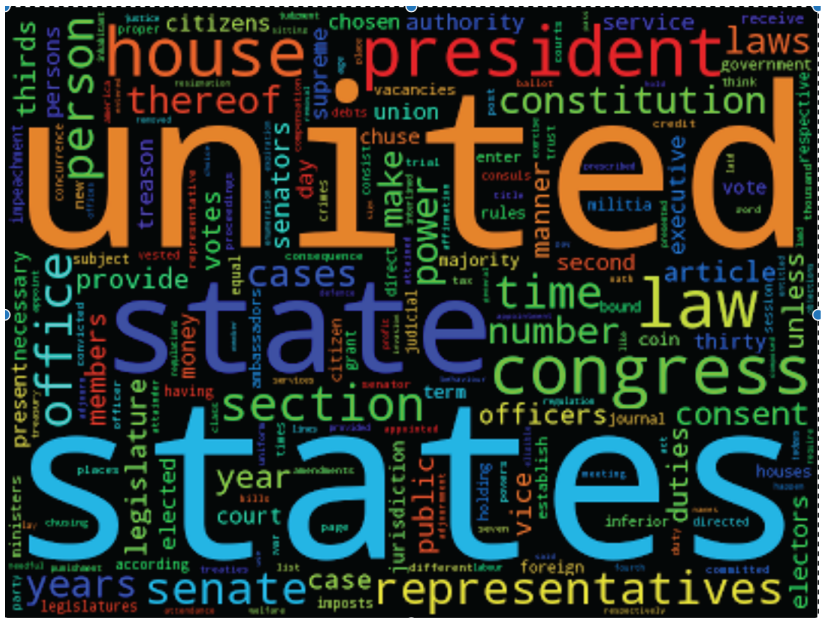
Parallel Axis Plot



Word-Clouds

Very useful in **text analytics**

A **word cloud** shows words used in a text corpus (a collection of documents) with size of words proportional to their importance (e.g. TF-IDF)



Quite clear that the word cloud on left is for a collection of articles about US politics, political news, while that on the right seems a corpus of astronomy/astrophysics