Exploratory Data Analysis (EDA)

Outline

- Types of data
- Data objects and attributes
- Types, properties, and sources of datasets
- Data exploration
- Issues of data quality
- Handling data quality issues

Variable

- A variable or a feature is any characteristic, number, or quantity that can be measured or counted
- E.g.,
  - Age (21, 35, 62, ...)
  - Gender (male, female)
  - Income ($25000, $35000, $50000, ...)
  - House price ($450000, $980000, ...)
  - Country of birth (Qatar, Australia, Saudi, ...)
  - Eye colour (blue, brown, green, ...)
  - Vehicle make (Toyota, Kia, ...)

Variable Types

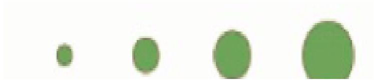
Type	Subtype	Examples
Categorical (Qualitative)	**Nominal**	Product type, name
	Ordinal	Size measured as small<medium<large
	Binary	Spam email (yes/no, true/false, 0/1)
	Date / Time	Job start date
Numerical (Quantitative)	**Discrete**	Number of students in a class
	Continuous	Height, weight

Understanding the type of variables is crucial for selecting appropriate statistical methods, visualization techniques, and ML algorithms

Categorical Variables

- Categorical data are strings that represent qualitative data
  - Often selected from a group of categories, also called labels
- **Nominal**, e.g., country of birth, gender, eye color, etc.

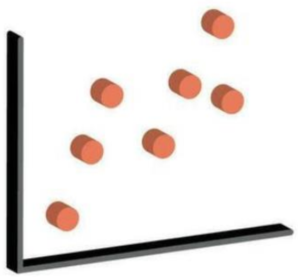
  – No inherent order or ranking
  – Operators applicable: **=, ≠**
  – 1:1 transformation permissible, e.g. ID: 974 ⇒ Qatar
- **Ordinal**, e.g. grade (A, B, C, D, F), degree (bachelor, master, PhD), height (tall, medium, short), etc.



  – Represent categories that can be meaningfully ordered
  – Operator applicable: **=, ≠, <, >, ≥, ≤**
  – Order-preserving transformation permitted,
  • e.g. height (tall, medium, short) to (1, 2, 3)

Numerical Variables

- **Discrete**



  - Whole numbers (counts) typically integers
  - E.g., The number of cars in a parking lot, the number of students in a class, or the count of items in a basket.

- **Continuous**

  - Measurable numeric variable that may contain any value within a range
  - Typically represented decimal numbers and fractions
  - E.g., Height, weight, temperature, or distance

Data Objects and Attributes

- **Data object:** (also known as record, sample, or entity) individual object/event
  - Characterized by its recorded values on a fixed set of features

- **Feature or attribute:** (also known as variable, field, or characteristic) a specific property or characteristic of the data object
  - **Raw Features:**
    - **Collected or measured** value of an attribute according to an appropriate measurement scale
  - **Derived Features**
    - Constructed from data in one or more raw features

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Derived Features

- **Aggregates:** defined over a group or period, e.g., count, sum, average, minimum, or maximum of the values
- **Flags:** indicate presence or absence of some characteristic within a dataset, e.g., a flag indicating whether or not a bank account has ever been overdrawn
- **Ratios:** capture relationship between two or more raw data values, e.g., a ratio between a loan applicant's salary and the amount for which they are requesting
- **Mappings:** convert continuous features into categorical features, e.g., map the salary values to low, medium, and high
- **Others:** no restrictions to the ways in which we can combine data to make derived features, e.g., use satellite photos to count the number of cars in the parking lots and use this as a proxy measure of activity within a competitor's stores!

Goals for Derived Features

- To **improve** the accuracy and performance of machine learning models by transforming the raw data into a more meaningful representation that can better capture the underlying relationships in the data

- To help to **reduce** the dimensionality of a dataset and make it easier to visualize and understand the relationships between variables

Types of datasets

Age Group	Own Car	Income Band	Class
young	yes	low	risky
young	no	low	risky
middle aged	yes	middle	risky
middle aged	no	high	safe
middle aged	yes	low	risky
young	yes	high	risky
middle aged	no	low	safe
retired	yes	middle	safe
retired	no	middle	safe
retired	yes	high	safe

Relational Table

No.	studentID Numeric	Homework1 Numeric	Homework2 Numeric	Homework3 Numeric	Final Exam Numeric
1	1.0		94.0	34.0	42.0
2	2.0	35.0	94.0	85.0	45.0
3	3.0	31.0	46.0	22.0	48.0
4	4.0	46.0	90.0	60.0	50.0
5	5.0	52.0	94.0	49.0	50.0
6	6.0	58.0	94.0	30.0	51.0
7	7.0	47.0	90.0		52.0
8	8.0	37.0	94.0	25.0	52.0
9	9.0	35.0	94.0	45.0	54.0
10	10.0	57.0	94.0	100.0	54.0
11	11.0	51.0	94.0	5.0	54.0
12	12.0	45.0	94.0	33.0	55.0
13	13.0	44.0	0.0	35.0	55.0
14	14.0	52.0	95.0	56.0	56.0
15	15.0	35.0	94.0		57.0
16	16.0	57.0	97.0	57.0	57.0
17	17.0	45.0	90.0	71.0	57.0
18	18.0	39.0	94.0	54.0	57.0
19	19.0	31.0	94.0	63.0	57.0
20	20.0	45.0	94.0		59.0
21	21.0	35.0	90.0	84.0	59.0
22	22.0	37.0	90.0	40.0	61.0
23	23.0	83.0	97.0	26.0	61.0
24	24.0	68.0	97.0	55.0	62.0
25	25.0	50.0	95.0	56.0	62.0
26	26.0	77.0	93.0		63.0
27	27.0	84.0	48.0	18.0	63.0

Data Matrix

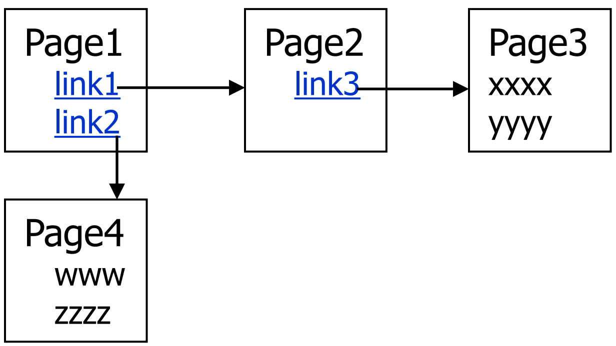
TID	Items
100	apple, milk, newspaper
200	apple, beef, milk, newspaper, potato
300	beef, potato
400	beef, noodles
500	beef, potato

Transaction Data

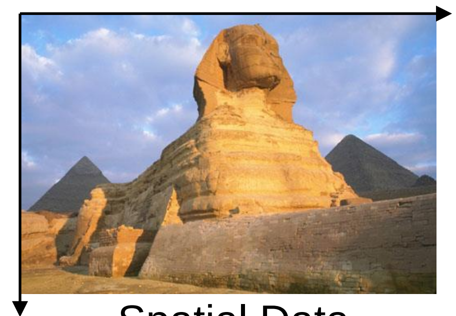
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Document-term Matrix

Types of data sets (cont.)



Web Structure



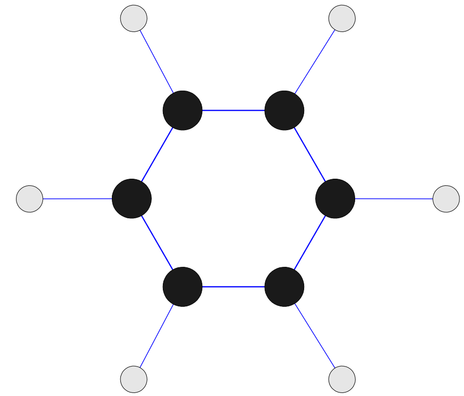
Spatial Data

GGTTCCGCCTTCAGCC
CCGCGCCCGCAGGG...

Data Sequence

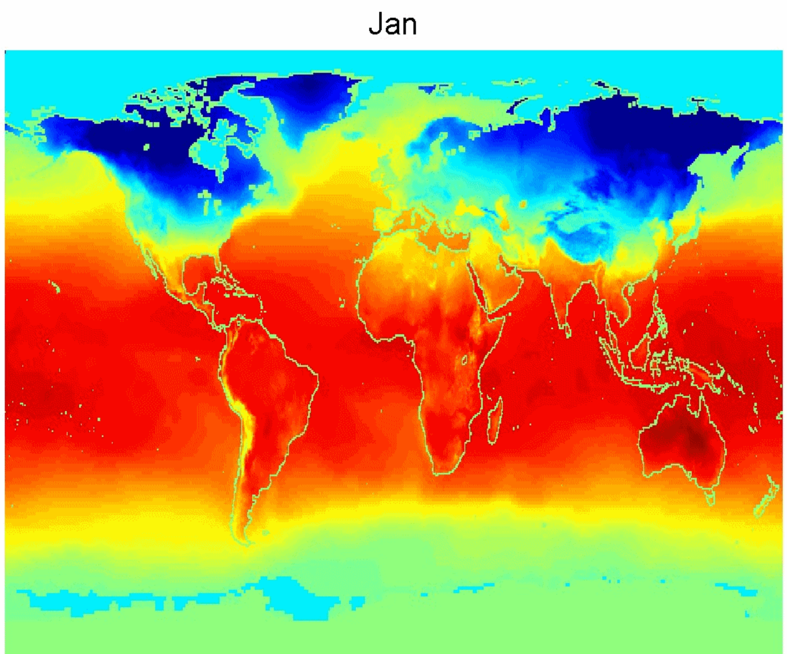
Types of data sets (cont.)

Chemical Data



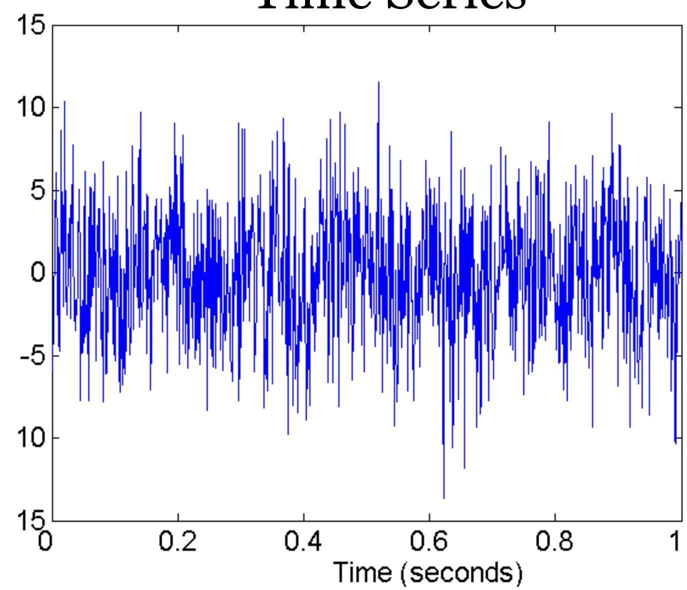
Benzene Molecule: C_6H_6

Spatio-Temporal Data



Average Monthly Temperature of land and ocean

Time Series



Data Matrix

- Data can often be represented or abstracted as an $n \times d$ *data matrix*, with *n* rows and *d* columns, given as

$$
\boldsymbol{D} = \left(\begin{array}{c|cccc}
 & X_1 & X_2 & \cdots & X_d \\
\hline
\boldsymbol{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\
\boldsymbol{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd}
\end{array}\right)
$$

- **Rows:** Also called *instances*, *examples*, *records*, *transactions*, *objects*, *points*, *feature-vectors*, etc. Given as a *d* -tuple

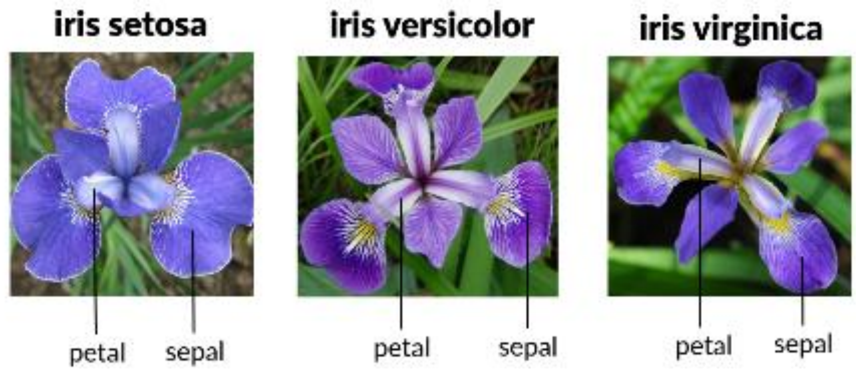
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

- **Columns:** Also called *attributes*, *properties*, *features*, *dimensions*, *variables*, *fields*, etc. Given as an *n*-tuple

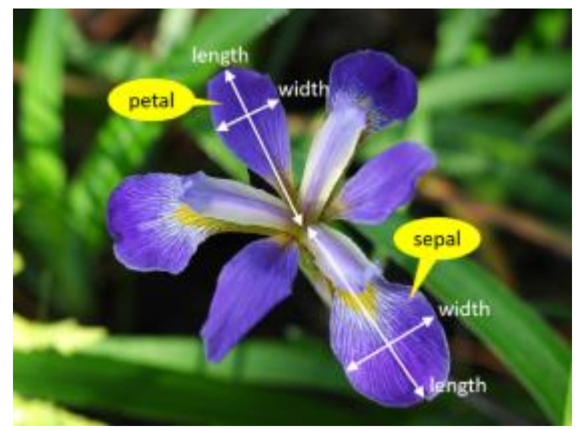
$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Iris Dataset Extract

Data to quantify
the morphologic variation
of *Iris* flowers
Wikipedia

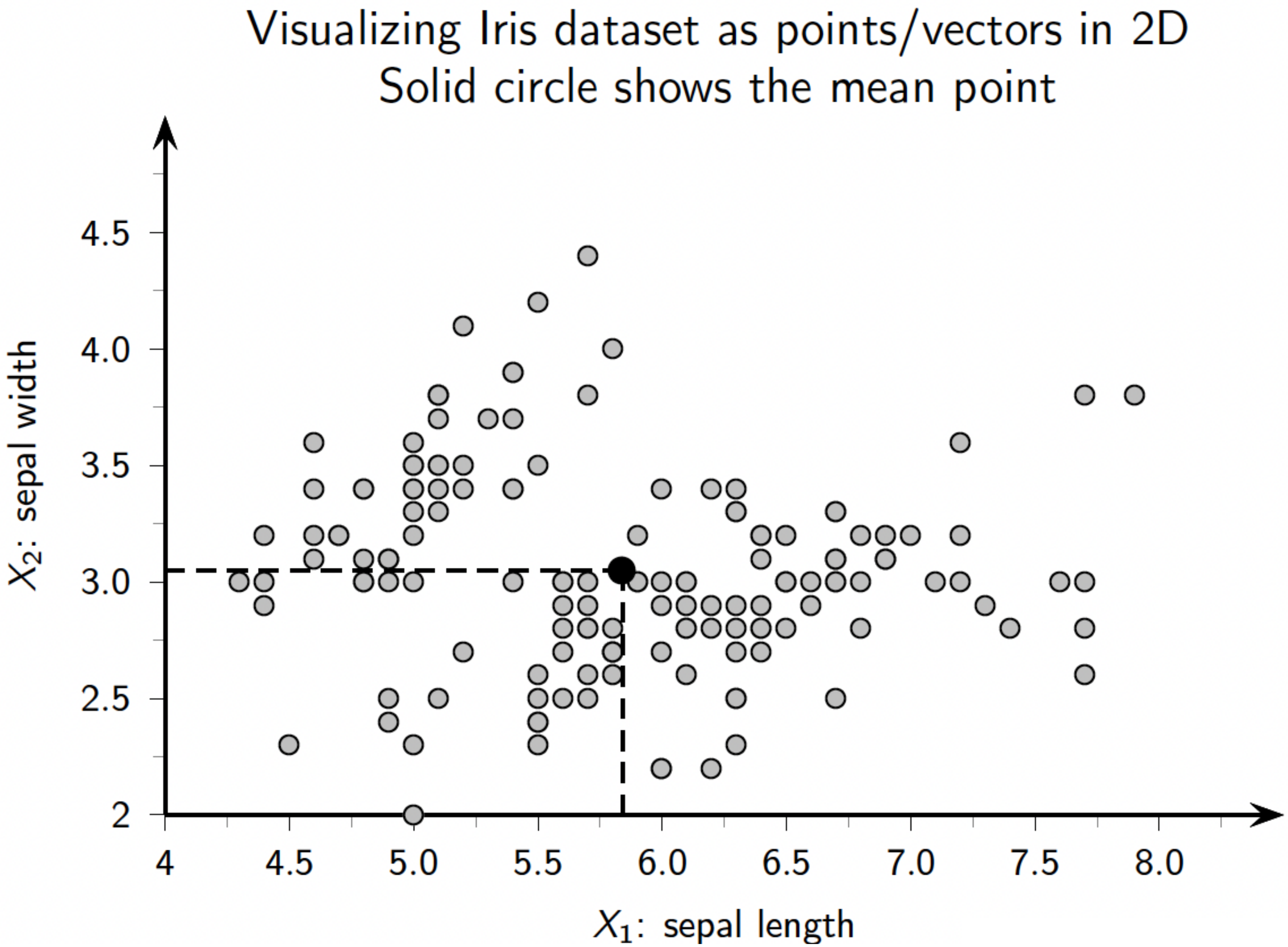


	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa



Scatterplot: 2D Iris Dataset

sepal length **versus** sepal width



Dataset Properties

Size:

Measured in terms of the total number of records or total number of bytes, e.g. Small (MB), medium (GB) and large (TB)

Dimensionality:

Number of attributes

Sparsity:

- Values are skewed to some extreme or sub-ranges
- Asymmetric values (some are more important than others)

Resolution:

- Right level of data details
- Related to the intended purpose

Data Sources

- **Public data**
  - Data hubs https://www.kaggle.com/datasets
  - Machine learning challenges
  - Data conferences
  - Many others…
- **Enterprise/Organisational data warehouse**
  - An organisational database for decision making
  - A central data repository separate from operational systems
  - Equipped with data analysis and reporting tools
- **Your own generated/collected data**

Data Exploration

- **Purpose:**
  - Better understanding of the characteristics of data
  - Better decision over data pre-processing tasks

- **Categories of data exploration techniques**
  - **Summary statistics:** using a small set of descriptors to describe the characteristics of a large data set

  - **Data visualisation:** using graphical or tabular forms to reveal hidden data patterns

Summary Statistics - Central Tendency

- Mean and Median for continuous attributes:

  - **Mean**

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - **Median** (Middle value if odd number of values, or average of the middle two values otherwise)

  Median is a better indication of "average" when data distribution is skewed, or outliers are present

  - Trimmed Mean and Median (after trimming top and bottom p%)

Summary Statistics - Central Tendency

- **Mode** for categorical attributes:
  - Frequency counts of values that a feature takes
  - Proportion: Frequency count for a value divided by the total sample size
  - **Mode**: the most frequently occurred value

Summary Statistics - Measures of Spread

- Range

$$range(x) = \max(x) - \min(x)$$

- Variance (σ^2)

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

- Standard Deviation (σ)

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

- Percentiles continuous attributes:
  - Given an attribute *x* and an integer *p* ($0 \le p \le 100$), the percentile x_p is a value of x such that *p%* observed values of *x* are less than x_p. Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$

Summary Statistics using Pandas

```
df.describe()
```

```
df[['DepTime', 'DepDelay', 'ArrTime',
'ArrDelay']].agg(['mean', 'min', 'max'])
```

```
price_mean = df['price'].mean()
```

Measures of Spread - Example

```
print( "Mean: {:.2f}".format(df['price'].mean()))
print( "Median: {:.2f}".format(df['price'].median()))
print( "Std: {:.2f}".format(df['price'].std()))
print( "Var: {:.2f}".format(df['price'].var()))
print( "Quantiles: \n",df['price'].quantile([0.25,0.5,0.75]))
```

Multivariate Summary Statistics

- Measures relationship between pairs of continues features
  - Covariance $\sigma_{xy} = \text{covariance}(x, y) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})$
    - a measure of the linear relations
    - measures the extent to which the variables change together.
  - The covariance matrix is a d x d (square) symmetric matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

  - Correlation (Pearson's Correlation Coefficient) $\rho_{x,y} = \text{correlation}(x, y) = \frac{\text{covariance}(x, y)}{\sigma_x \sigma_y}$
    - a measure of the linear relations
    - Between −1 and +1
    - If >0 or < 0, positively/negatively correlated (x's values increase/decrease as y's).
    - The closer to +1 or -1, the stronger correlation.
    - If = 0: independent.
  - The correlation matrix is a d x d (square) symmetric matrix

$$\begin{pmatrix} \rho_1^2 & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{21} & \rho_2^2 & \cdots & \rho_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{d1} & \rho_{d2} & \cdots & \rho_d^2 \end{pmatrix}$$

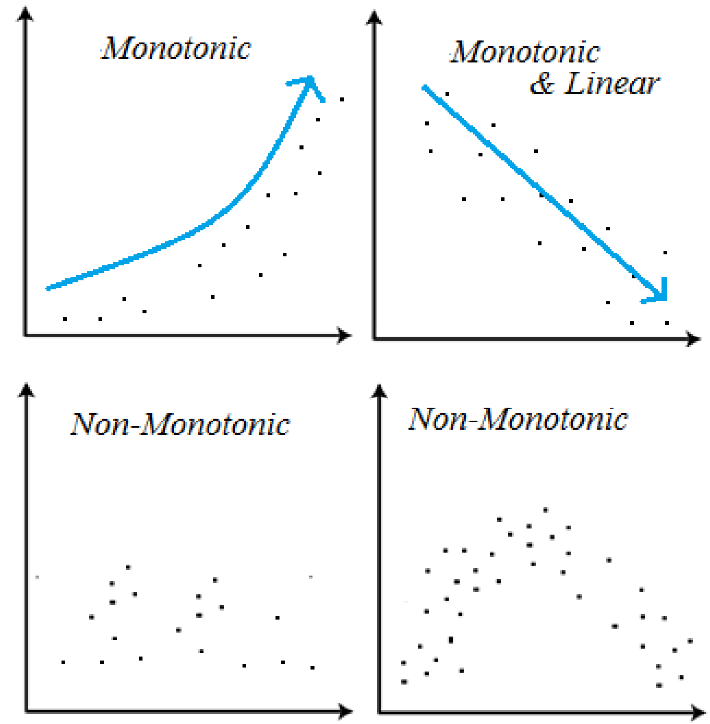
Pearson's correlation coefficient

- Pearson's correlation coefficient is a parametric measure of the linear relationship between two continuous variables. As a parametric method, it makes certain assumptions about the data, including:
  - Linearity: there is a linear relationship between the two variables. If the relationship between the variables is not linear, Pearson's correlation may not accurately reflect the relationship.
  - Normality: the data is normally distributed. This means that the distribution of the residuals (the difference between the values) should follow a normal distribution.
  - Independence: the observations are independent of one another. This means that the value of one observation does not influence the value of another observation.

If these assumptions are not met, Pearson's correlation may not accurately reflect the relationship between the variables. In these cases, non-parametric methods, such as Spearman's rank correlation, may be more appropriate.

Spearman's Rank Correlation

- Spearman's Rank Correlation
  - non-parametric: does not assume a specific distribution of the data
  - measure of the monotonic relations: the variables tend to move in the same direction, but not necessarily at a constant
  - measure the degree of correlation between two variables
  - calculated based on the ranks of the data points instead of the actual values.



- Example:

Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

Students	Maths	Rank	Science	Rank	d	d square
A	35	3	24	5	2	4
B	20	5	35	4	1	1
C	49	1	39	3	2	4
D	44	2	48	1	1	1
E	30	4	45	2	2	4
						14

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

1 - (6 * 14) / 5(25 - 1) = 0.3

The Spearman’s Rank Correlation for the given data is 0.3. The value is near 0, which means that there is a weak correlation between the two ranks.

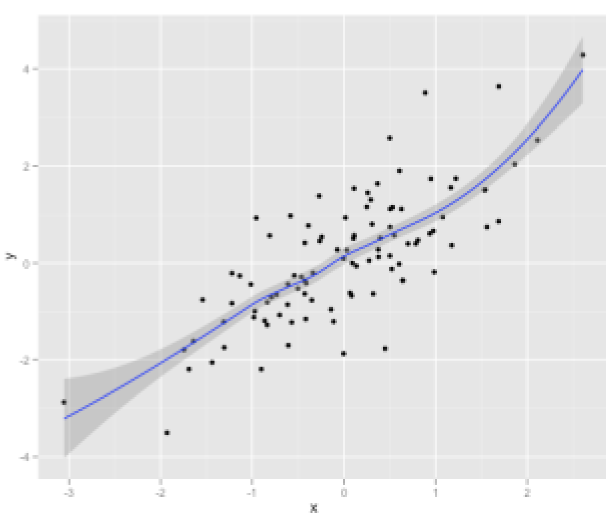
Data Exploration (Visualization)

Summary statistics give us some sense of the data:

- Mean vs. Median.
- Standard deviation
- Quartiles, Min/Max
- Correlations between variables.

```
Summary(data)
       x                    y
 Min.   :-3.05439   Min.   :-3.50179
 1st Qu.:-0.61055   1st Qu.:-0.75968
 Median : 0.04666   Median : 0.07340
 Mean   :-0.01105   Mean   : 0.09383
 3rd Qu.: 0.56067   3rd Qu.: 0.88114
 Max.   : 2.60614   Max.   : 4.28693
```

Why Visualize?



Visualization gives us a more holistic sense

Anscombe's Quartet

4 data sets, characterized by the following. Are they the same, or are they different?

Property	Values
Mean of x in each case	9
Exact variance of x in each case	11
Exact mean of y in each case	7.5 (to 2 d.p)
Variance of Y in each case	4.13 (to 2 d.p)
Correlations between x and y in each case	0.816
Linear regression line in each case	Y = 3.00 + 0.500x (to 2 d.p and 3 d.p resp.)

i

X	y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

ii

x	y
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

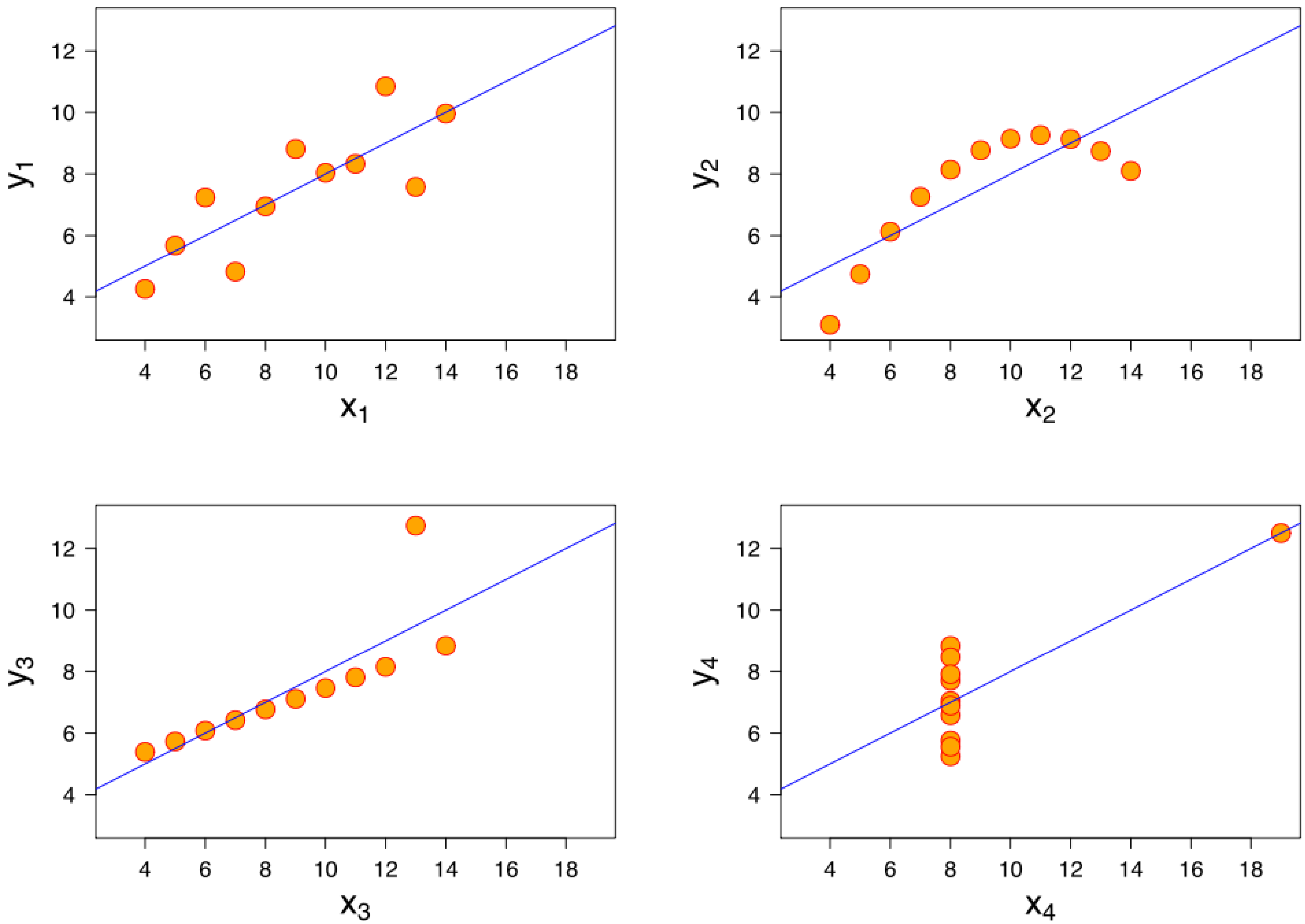
iii

x	y
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

iv

x	y
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

Moral: Visualize Before Analyzing!



Visualizing Your Data

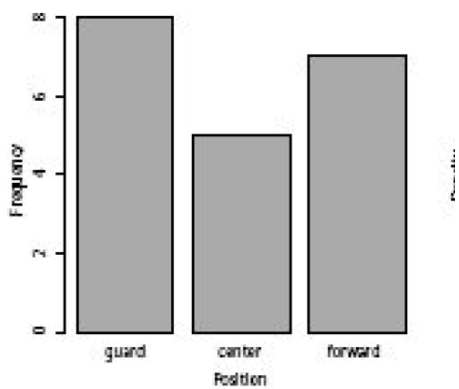
- Examining the distribution of a single variable
- Analyzing a single variable over time
- Analyzing the relationship between two variables
- Establishing multiple pair wise relationships between variables

Data visualization for a single feature

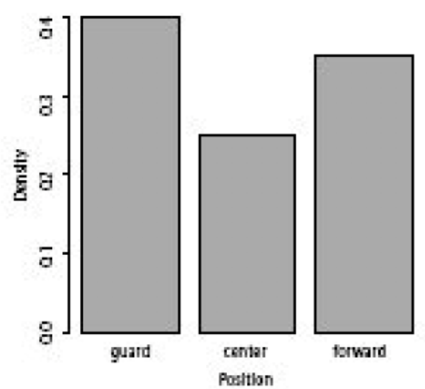
- Bar plot

A dataset showing the positions and monthly training expenses of a basketball team.

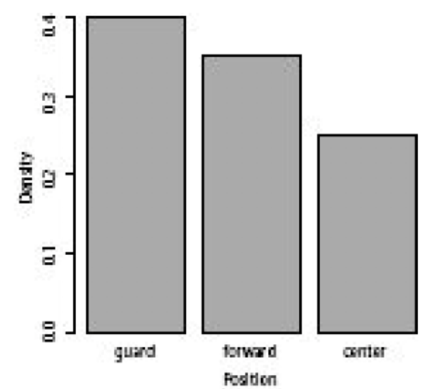
ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41



Frequency bar plot for
for the POSITION feature



Density bar plot.
(Probability distribution)

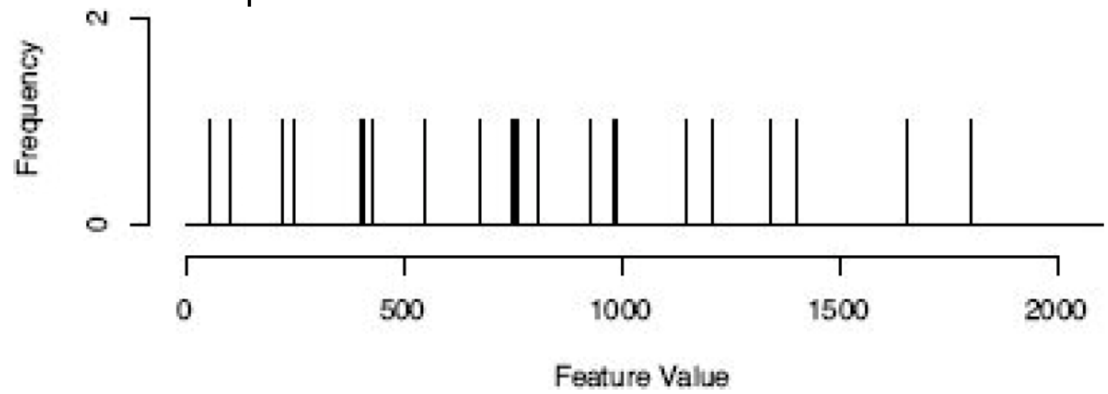


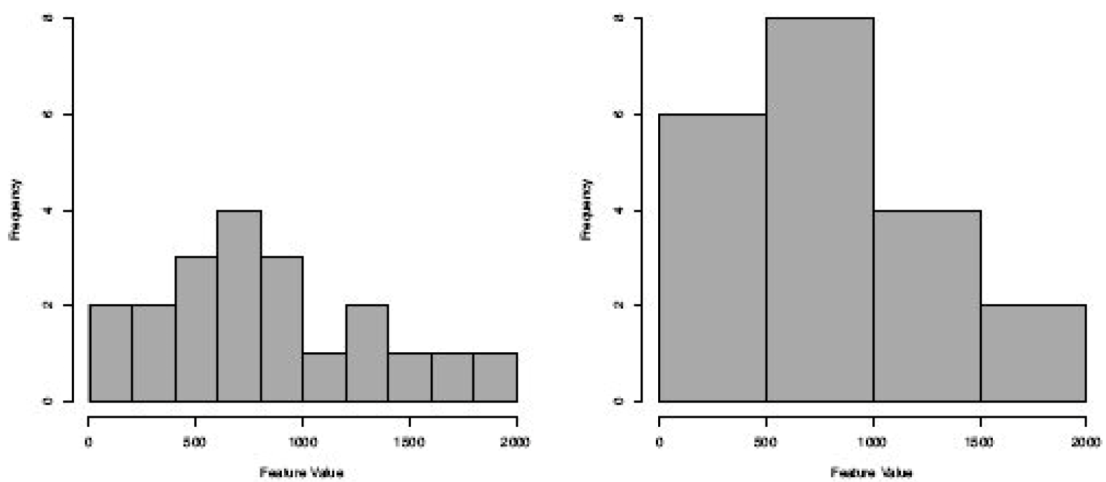
Order density bar plot.

Data visualization for a single feature

- Histogram,

Bar plot of the continuous TRAINING EXPENSES feature

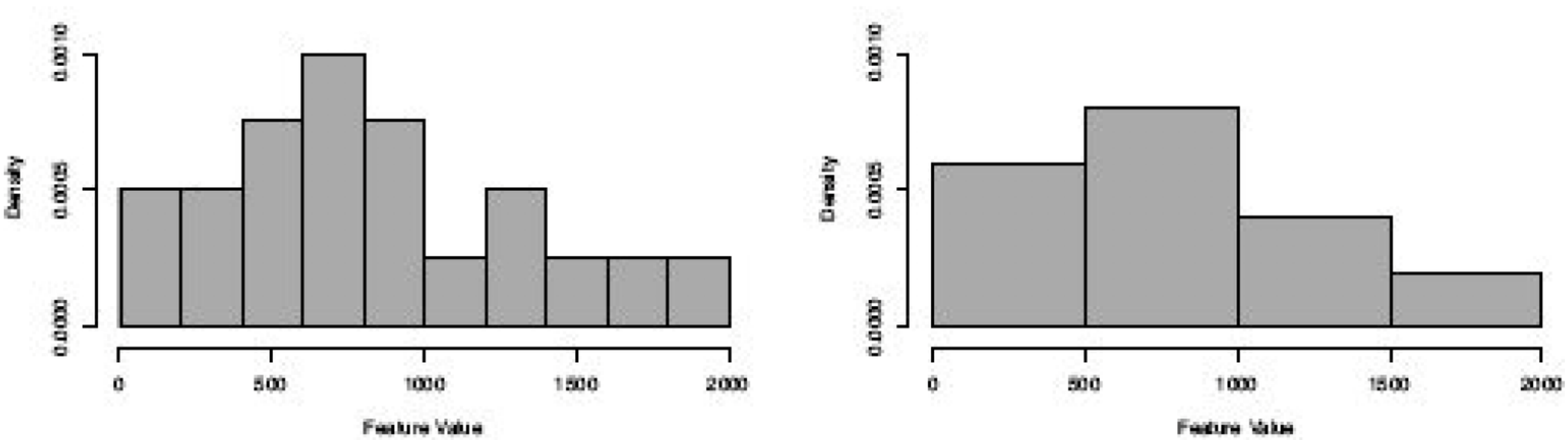




Frequency histograms (200/500-unit intervals) for the continuous TRAINING EXPENSES feature

Data visualization for a single feature

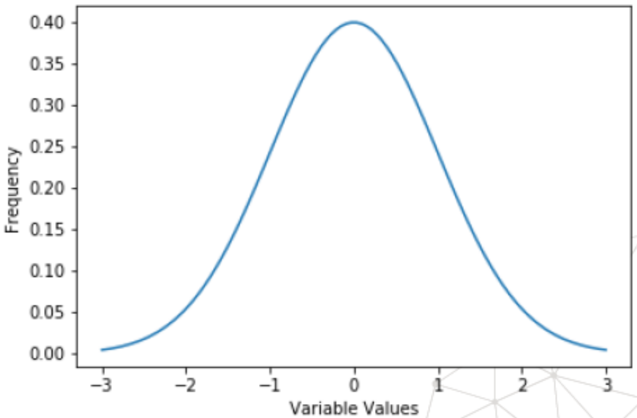
- Histogram to probability distribution
  - divide the count for each interval by the total number of observations in the dataset multiplied by the width of the interval



Density histograms (200/500-unit intervals) for the continuous TRAINING EXPENSES feature

Normal Distribution

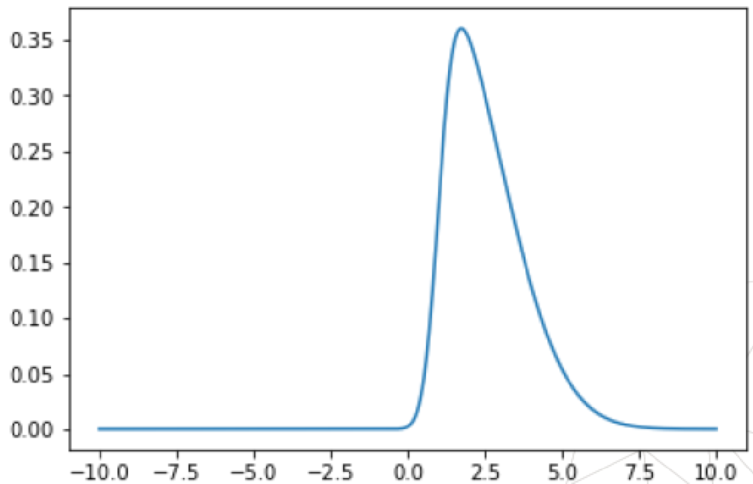
- Many natural phenomena follow a normal distribution
  - Height, blood pressure, etc.



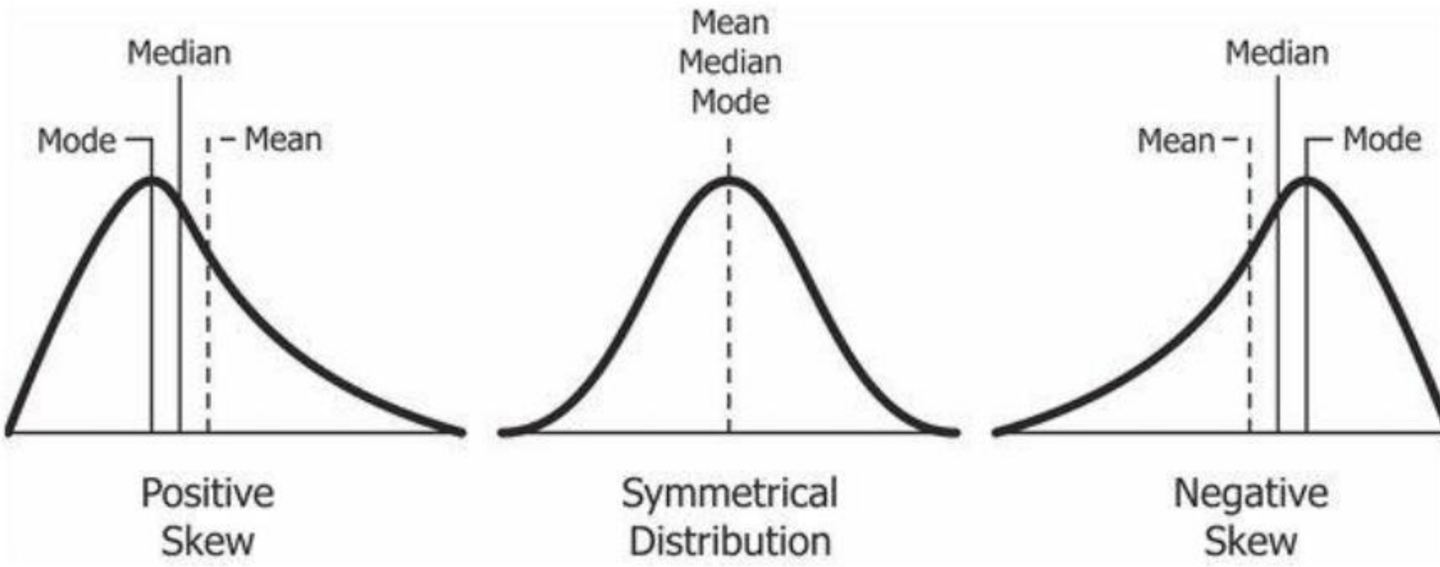
- Symmetric:
  - Most of the observations occur around the central peak
  - Probabilities for values further away from the centre decrease equally in both directions.
  - Extreme values in both tails of the distribution are similarly unlikely.

Skewed distributions

- A distribution is skewed if one of its tails is longer than the other
- A left-skewed distribution has a long left tail. Also called negatively-skewed distributions.
- A right-skewed distribution shows a long right tail. Also called positive-skew distributions.



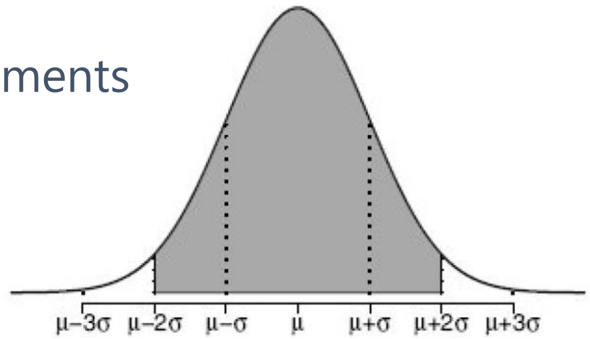
Gaussian vs Skewed distributions

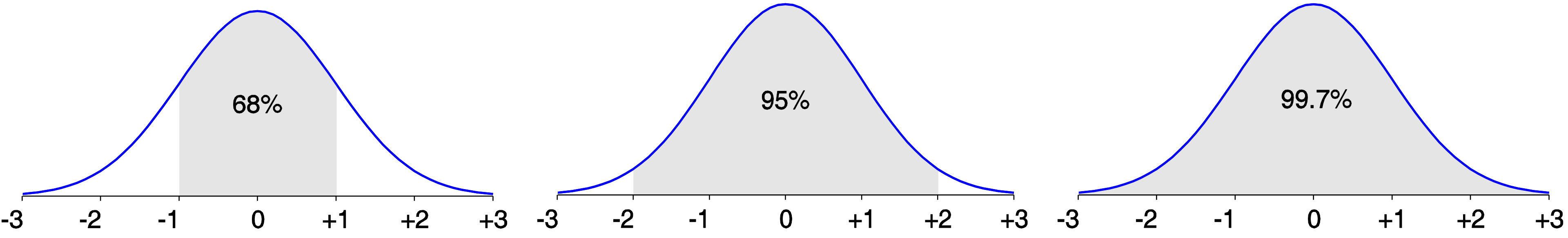


- In Normal distributions, the mean, median and mode are the same
- For skewed distributions, the mean is influenced by the tail

Spread Properties of Normal Distribution Curve

- The 68−95−99.7 rule:
  - From μ–σ to μ+σ: contains about 68% of the measurements
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

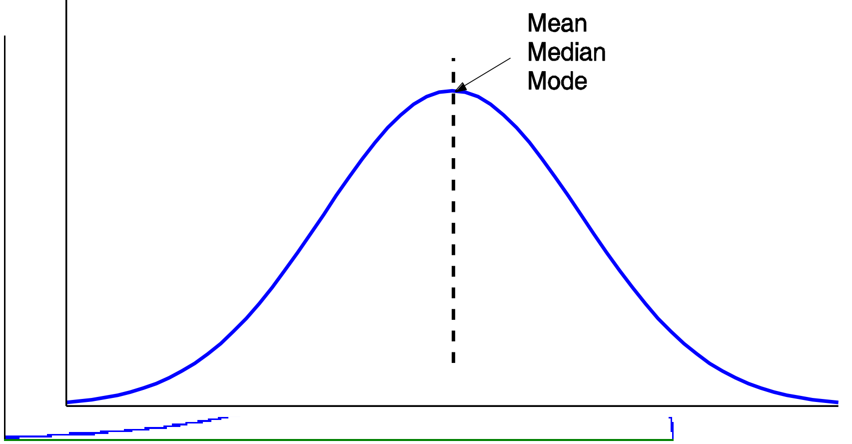




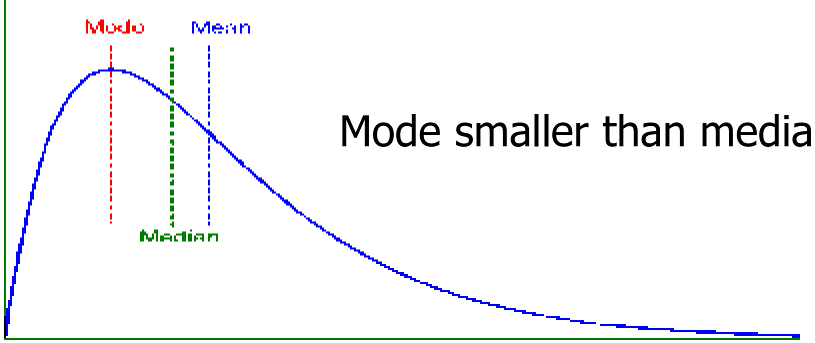
Symmetric vs. Skewed Data

- Central measures can indicate level of symmetry in data
- Median, mean and mode of symmetric, positively and negatively skewed data

Symmetric



Positive (right) Skew



Mode smaller than median

Negative (left) Skew

Mode larger than median

What are we looking for?

A sense of the data range

- If it's very wide, or very skewed, try computing the log

Outliers, anomalies

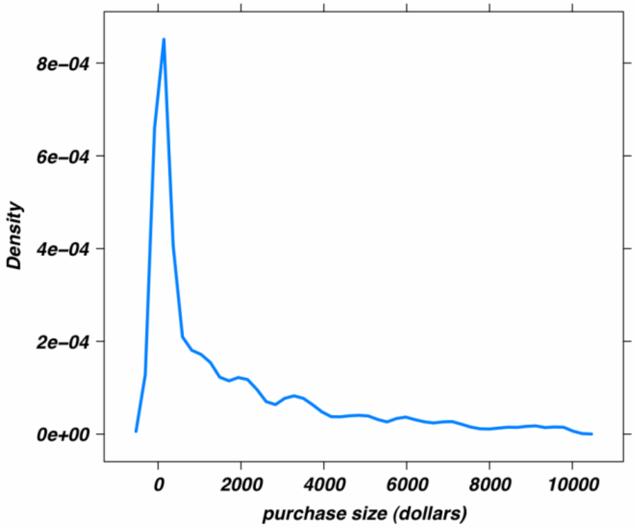
- Possibly evidence of dirty data

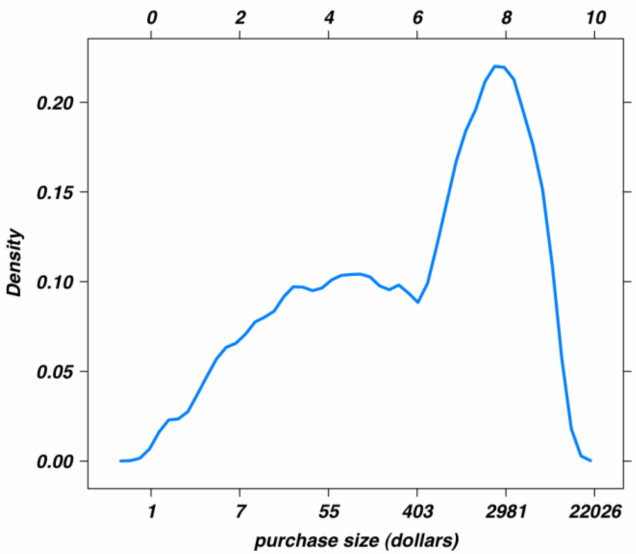
Shape of the Distribution

- Unimodal? Bimodal?
- Skewed to left or right?
- Approximately normal? Approximately lognormal?

Example - Distribution of purchase size ($)

- Range from 0 to > $10K, right skewed
- Typical of monetary data
- Plotting log of data gives better sense of distribution
- Two purchasing distributions
  - ~ $55
  - ~ $2900



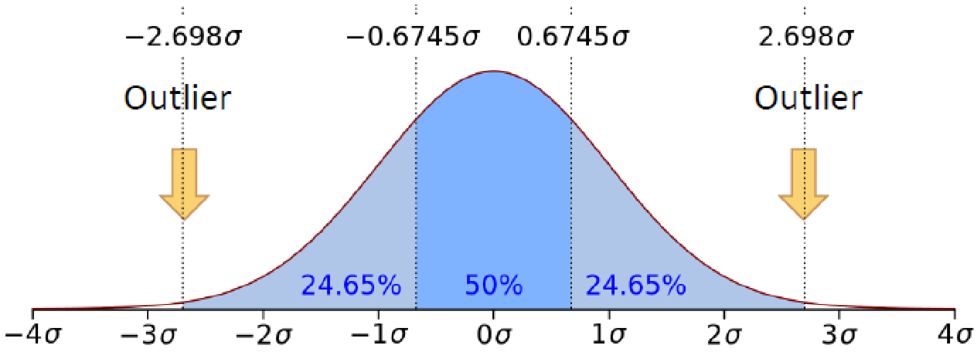


Ouliers

- An outlier is a data point which is significantly different from the remaining data.
- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” [D. Hawkins. Identification of Outliers, Chapman and Hall , 1980.]

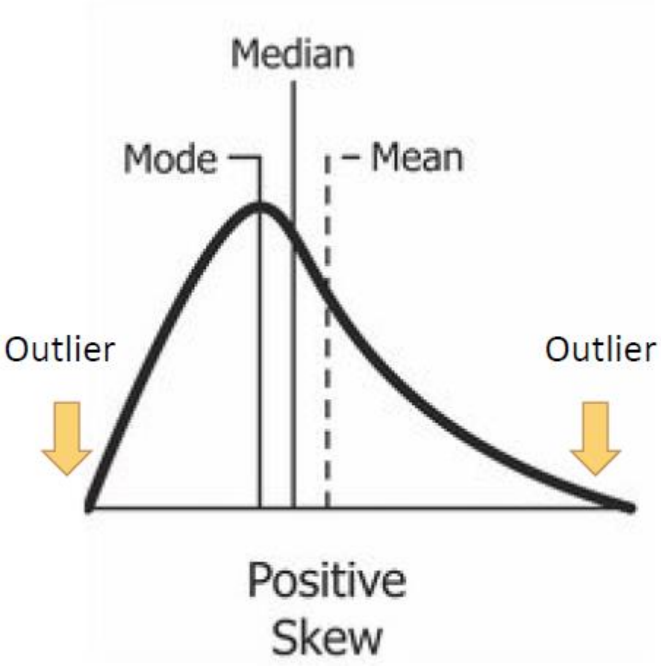
Detecting Outliers - Extreme Value Analysis

Normal Distribution



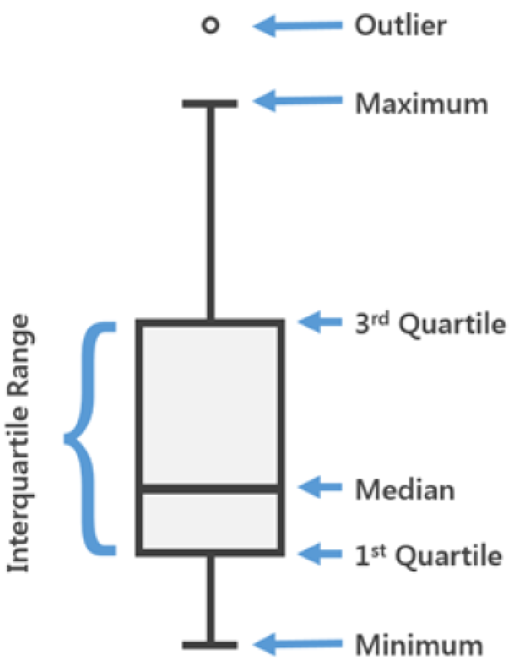
- ~99% of the observations of a normally distributed variable lie within the mean ± 3 × standard deviations.
- Values outside mean ± 3 × standard deviations are considered outliers

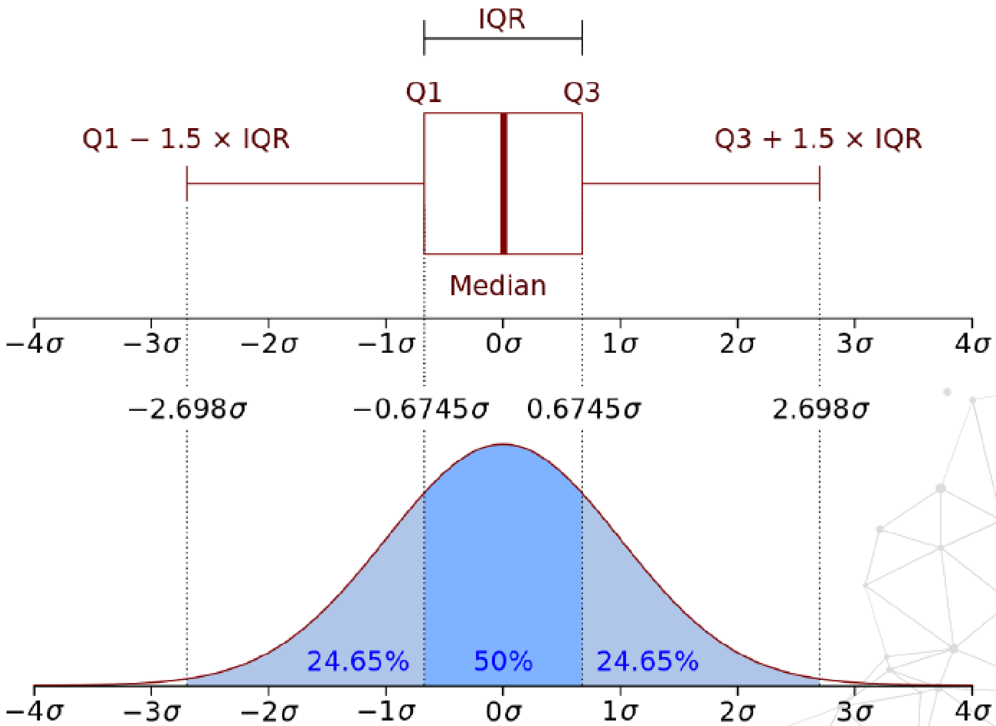
Skewed distributions



- The general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:
- IQR = 75^{th} Quantile – 25^{th} Quantile
- Upper limit = 75^{th} Quantile + IQR × 1.5
- Lower limit = 25^{th} Quantile - IQR × 1.5

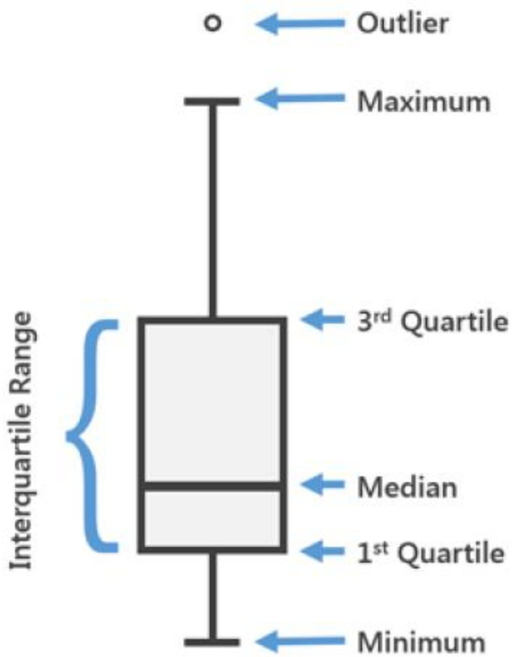
Visualizing outliers -Boxplots

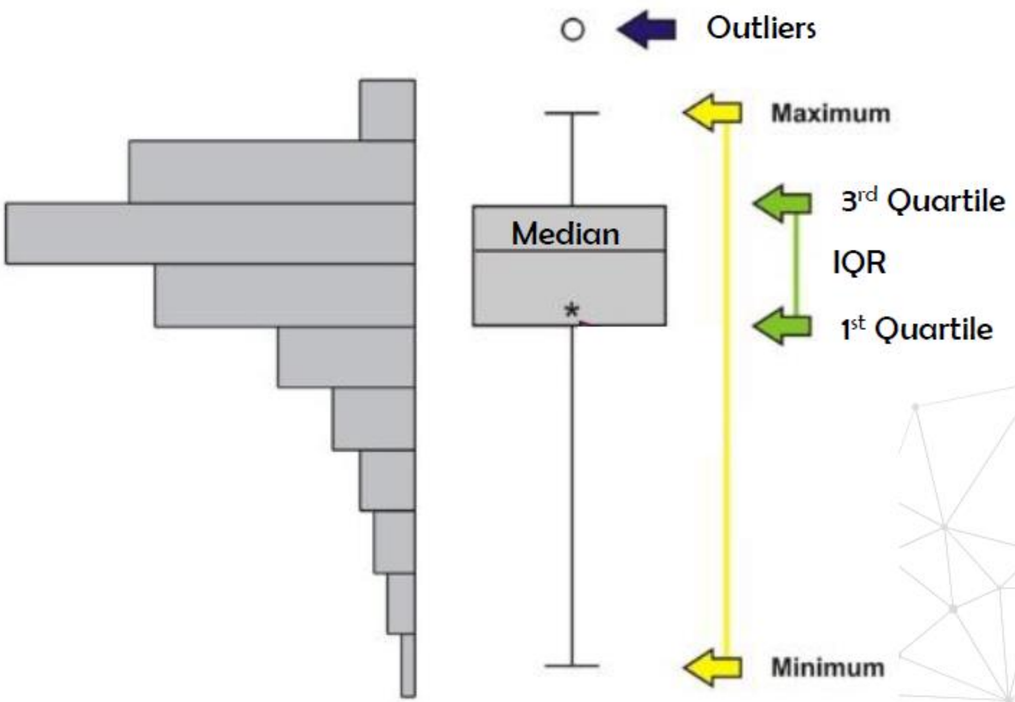




Images taken from [pro.arcgis.com]() and [wiki.commons]()

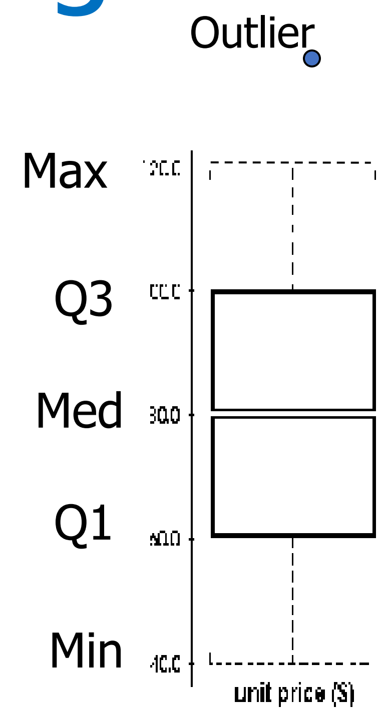
Visualizing outliers -Boxplots





Images taken from pro.arcgis.com and wiki.commons

Data visualization for a single feature



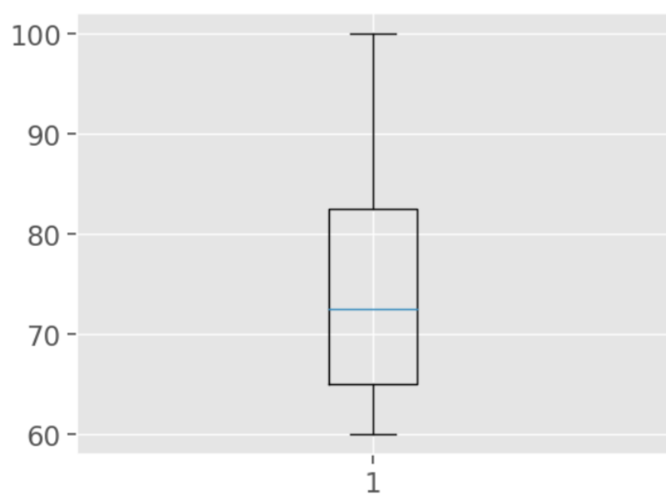
- Box plot
- Five-number summary of a distribution:

  Min, Q1, Med, Q3, Max
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend (usually) to Minimum and Maximum

Quartiles, outliers and boxplots

- Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
- Inter-quartile range: $IQR = Q_3 - Q_1$
- Five number summary: min, Q_1, Med, Q_3, max
- Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
- Outlier: usually, a value higher/lower than Q3/Q1 by 1.5 x IQR
- Example:

```
X = np.array([60,65,65,70,75,80,90,100])
plt.boxplot(X)
plt.show()
```



Grade:
65
70
80
90
65
100
60
75

60	Min
65	Q1
72.5	Med
85	Q3
100	MAX
20	IQR

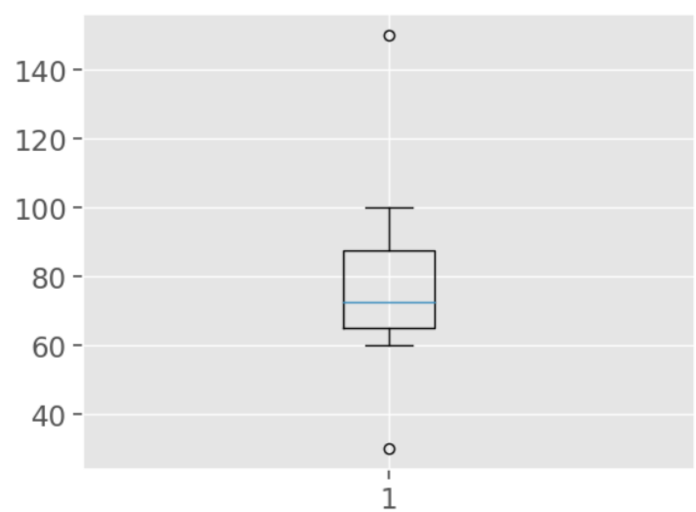
Outlier Grades:
Higher than:
85 + 1.5*20 = 115,

and Lower than:
65 – 1.5*20 = 35

Quartiles, outliers and boxplots

- Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
- Inter-quartile range: IQR = $Q_3 - Q_1$
- Five number summary: min, Q_1, Med, Q_3, max
- Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
- Outlier: usually, a value higher/lower than Q3/Q1 by 1.5 x IQR
- Example:

```
X = np.array([30,60,65,65,70,75,80,90,100,150])
plt.boxplot(X)
plt.show()
```



Grade:
65
70
80
90
65
100
60
75

60	Min
65	Q1
72.5	Med
85	Q3
100	MAX
20	IQR

Outlier Grades:
Higher than:
Q3 + 1.5 * IQR
85 + 1.5*20 = 115,

and Lower than:
Q1 - 1.5 * IQR
65 – 1.5*20 = 35

Analyzing a Single Variable over Time

What?

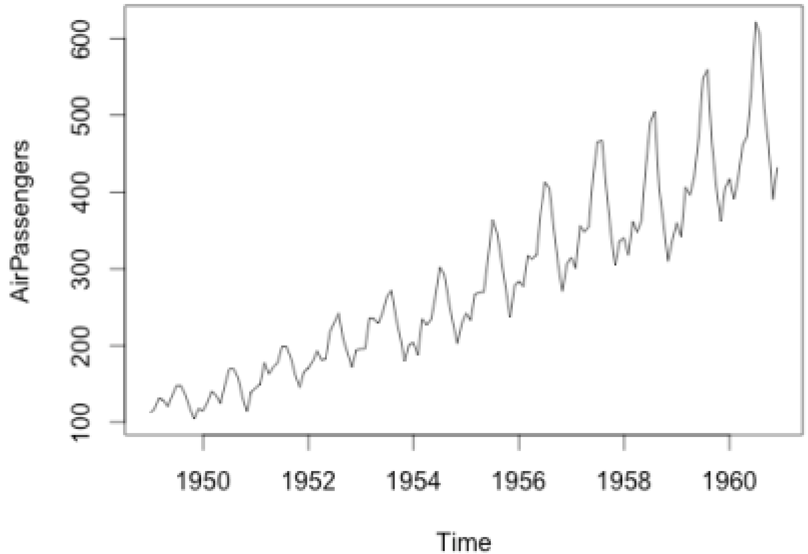
- Looking for ...
  - Data range
  - Trends
  - Seasonality

How?

  - Use time series plot

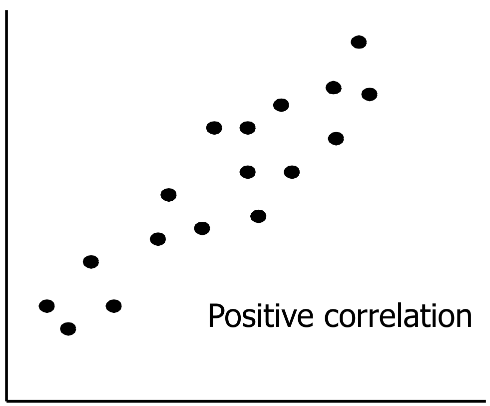
Example

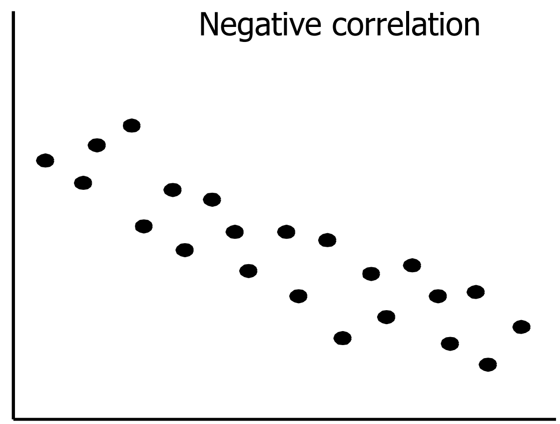
  - International air travel (1949-1960)
  - Upward trend: growth appears superlinear
  - Seasonality
    - Peak air travel around Nov. with smaller peaks near Mar. and June

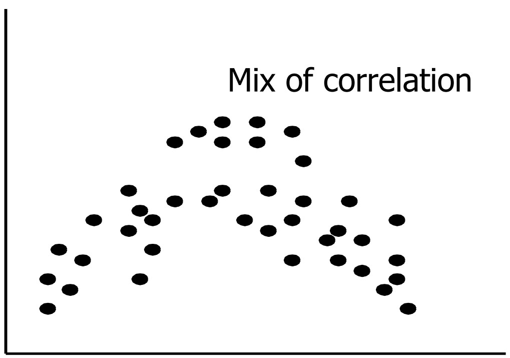


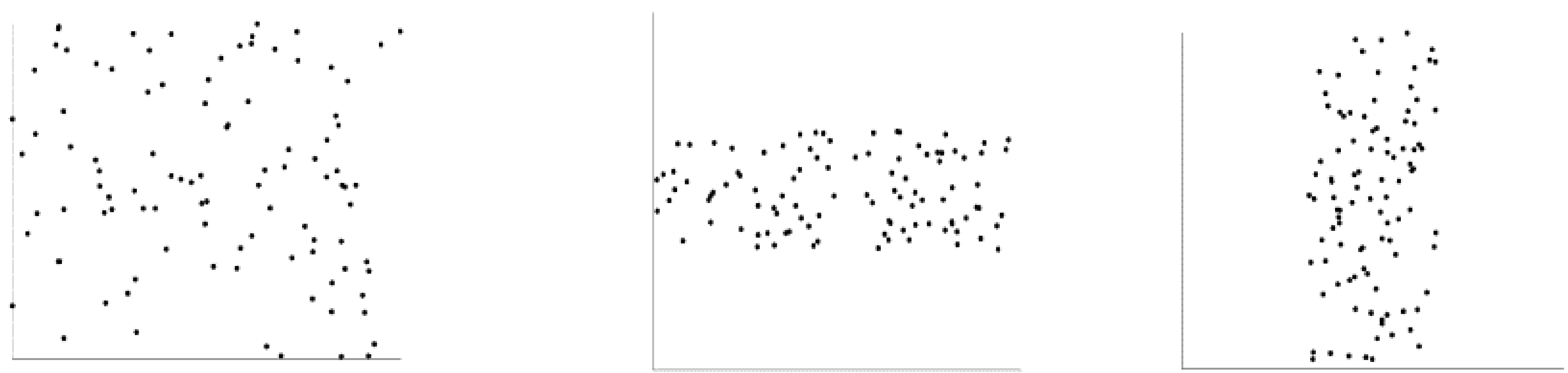
Analyzing the Relationship Between Two Variables

- **Scatter Plot:** Visualizing Pairs of Continuous Features









Not Correlated Data

Analyzing the Relationship Between Two Variables

- Visualizing Pairs of Continuous Features
  - Iris Characteristics: Strong linear relationship between petal length and width

Using Seaborn Library: statistical data visualization

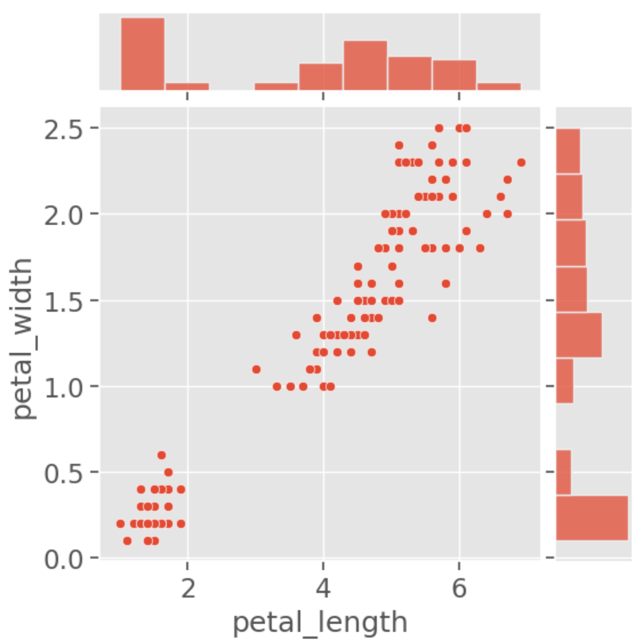
In [52]:
```
import seaborn as sns
sns.set_context("notebook", font_scale=1.5, rc={"lines.linewidth": 2.5})
iris = sns.load_dataset('iris') #Load iris dataset as a dataframe
```

In [53]:
```
iris.head()
```

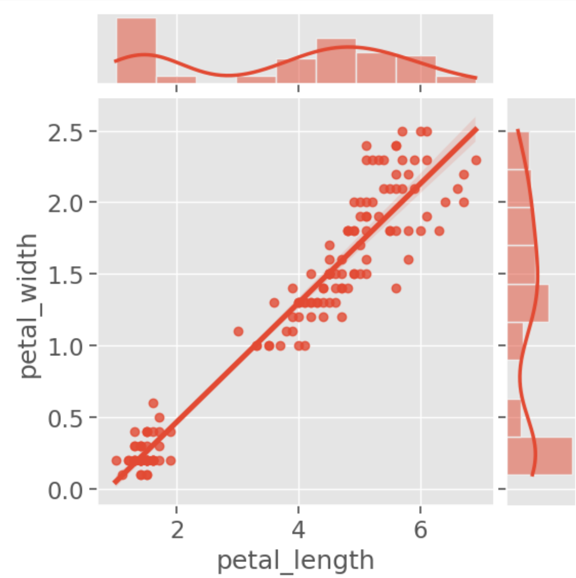
Out[53]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

```
sns.jointplot(x='petal_length', y='petal_width', data=iris, kind='scatter')
pass
```

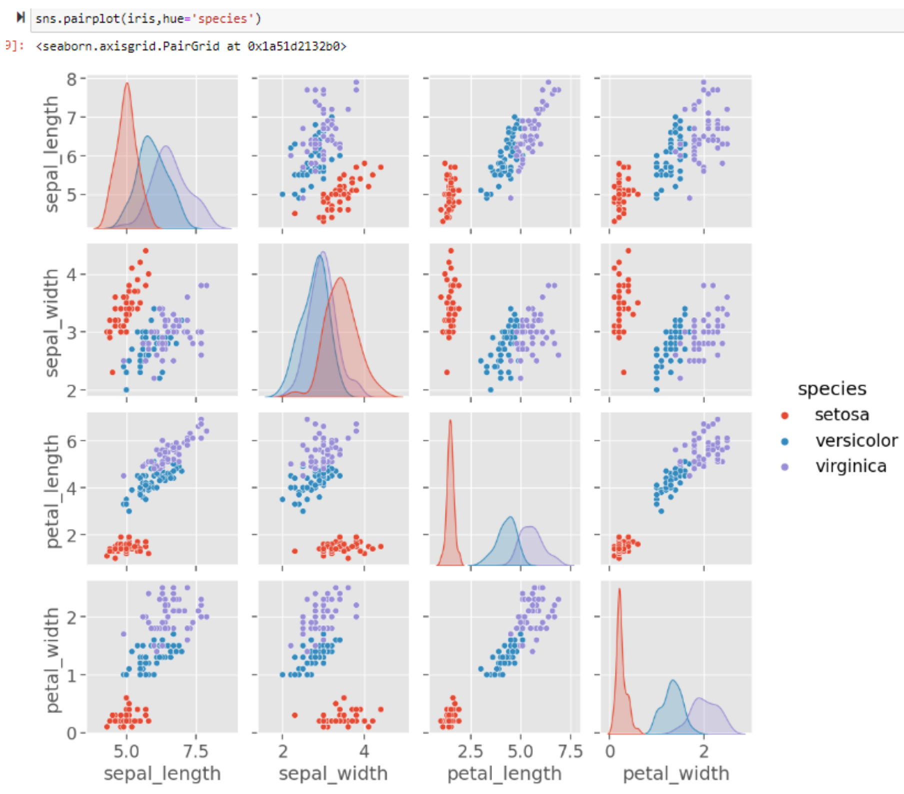



```
sns.jointplot(x='petal_length', y='petal_width', data=iris, kind='reg')
pass
```



Analyzing the Relationship Between Two Variables

- Visualizing Pairs of Continuous Features: Iris Characteristics
  - **Scatter Plot matrix**
    - Strong linear relationship between petal length and width
    - Petal dimensions discriminate species more strongly than sepal dimensions



Analyzing the Relationship Between Two Variables

- Visualizing Pairs of Continuous Features
  - **Correlation values**

```
cormat = iris.corr(method="pearson")
round(cormat,2)
```

]:

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.00	-0.12	0.87	0.82
sepal_width	-0.12	1.00	-0.43	-0.37
petal_length	0.87	-0.43	1.00	0.96
petal_width	0.82	-0.37	0.96	1.00

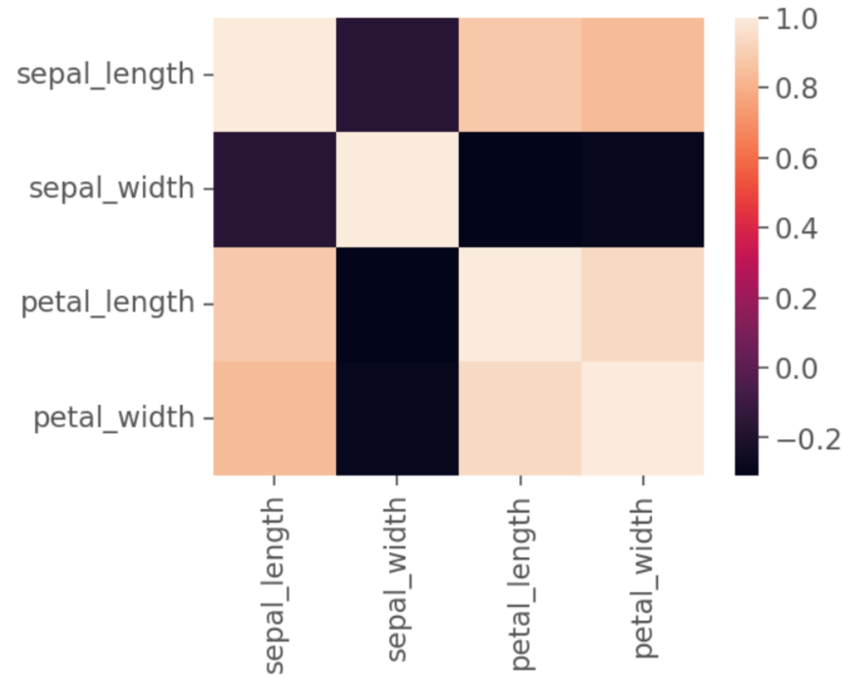
In [65]:
```
cormat = iris.corr(method="spearman")
round(cormat,2)
```

Out[65]:

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.00	-0.17	0.88	0.83
sepal_width	-0.17	1.00	-0.31	-0.29
petal_length	0.88	-0.31	1.00	0.94
petal_width	0.83	-0.29	0.94	1.00

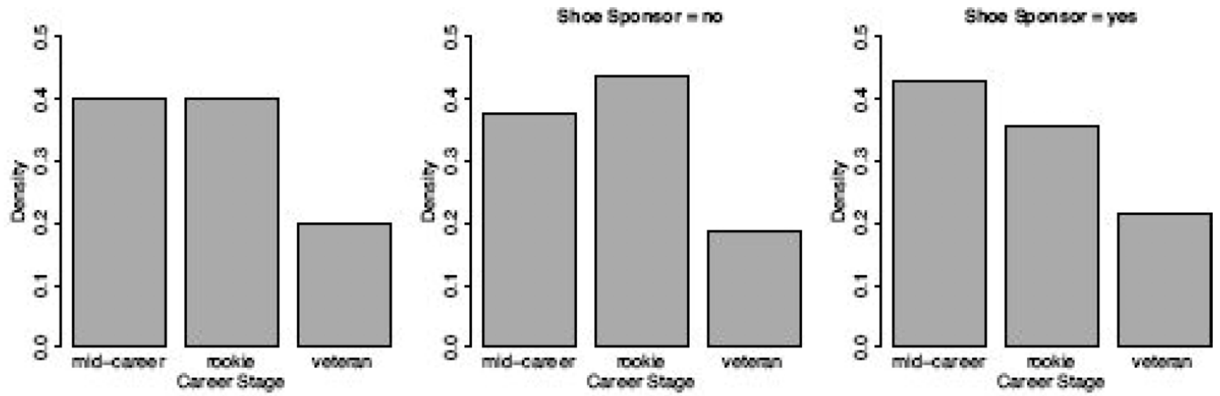
```
sns.heatmap(cormat)
```

7]: <AxesSubplot:>



Analyzing the Relationship Between Two Variables

- Visualizing Pairs of Categorical Features
  - **Collection of bar plots**

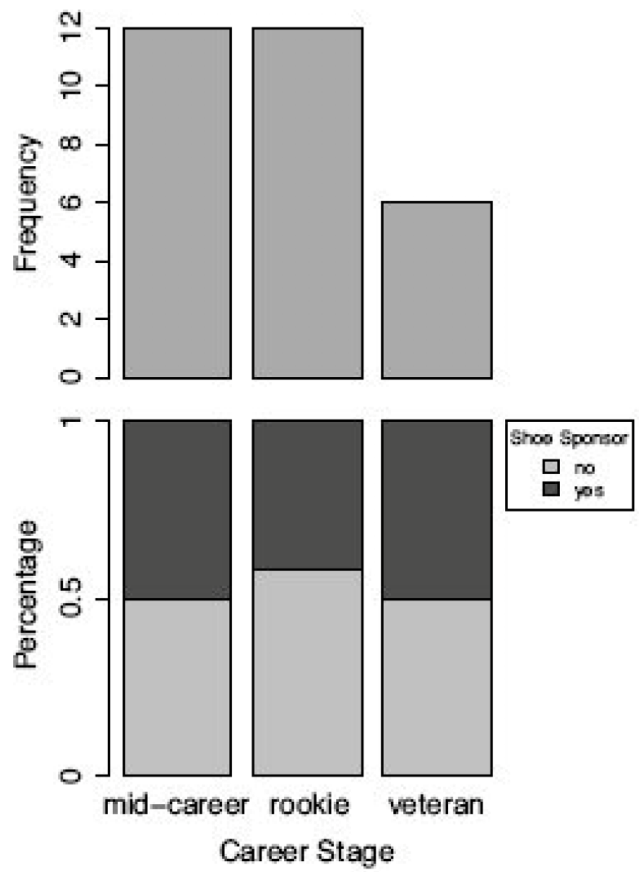


(a) Career Stage and Shoe Sponsor

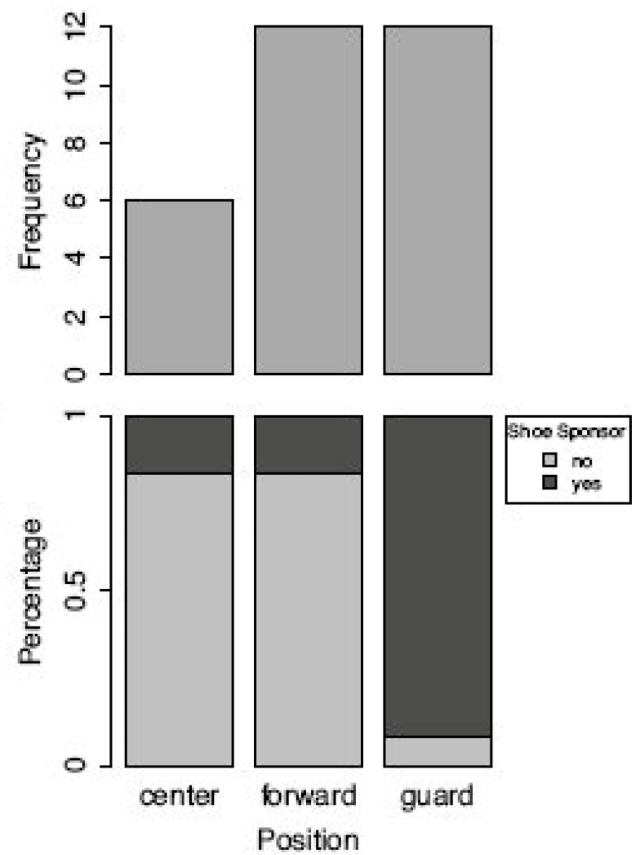
(b) Position and Shoe Sponsor

Analyzing the Relationship Between Two Variables

- Visualizing Pairs of Categorical Features
  - **Stacked bar plots**



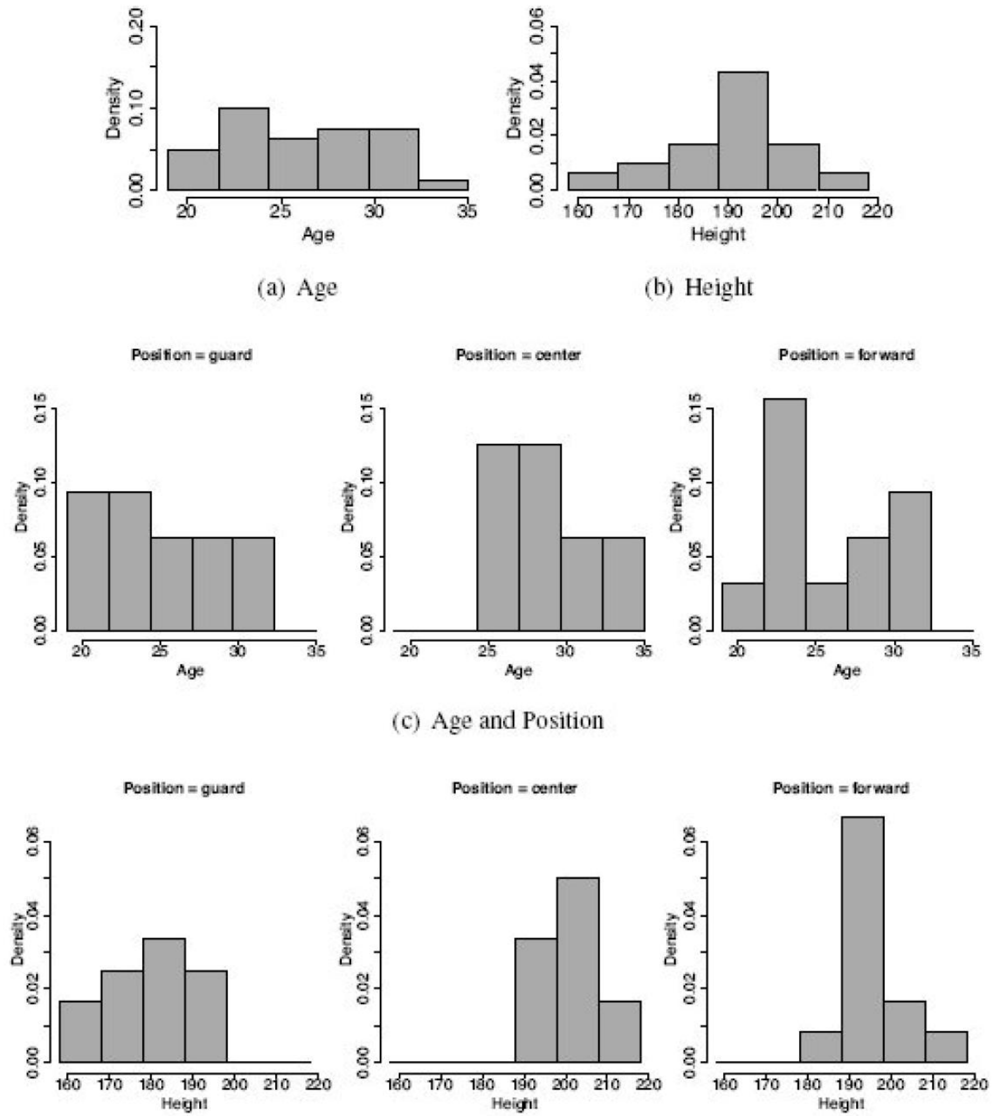
(a) Career Stage and Shoe Sponsor



(b) Position and Shoe Sponsor

Analyzing the Relationship Between Two Variables

- Visualizing a Categorical Feature and a Continuous Feature
  - **Collection of bar plots**



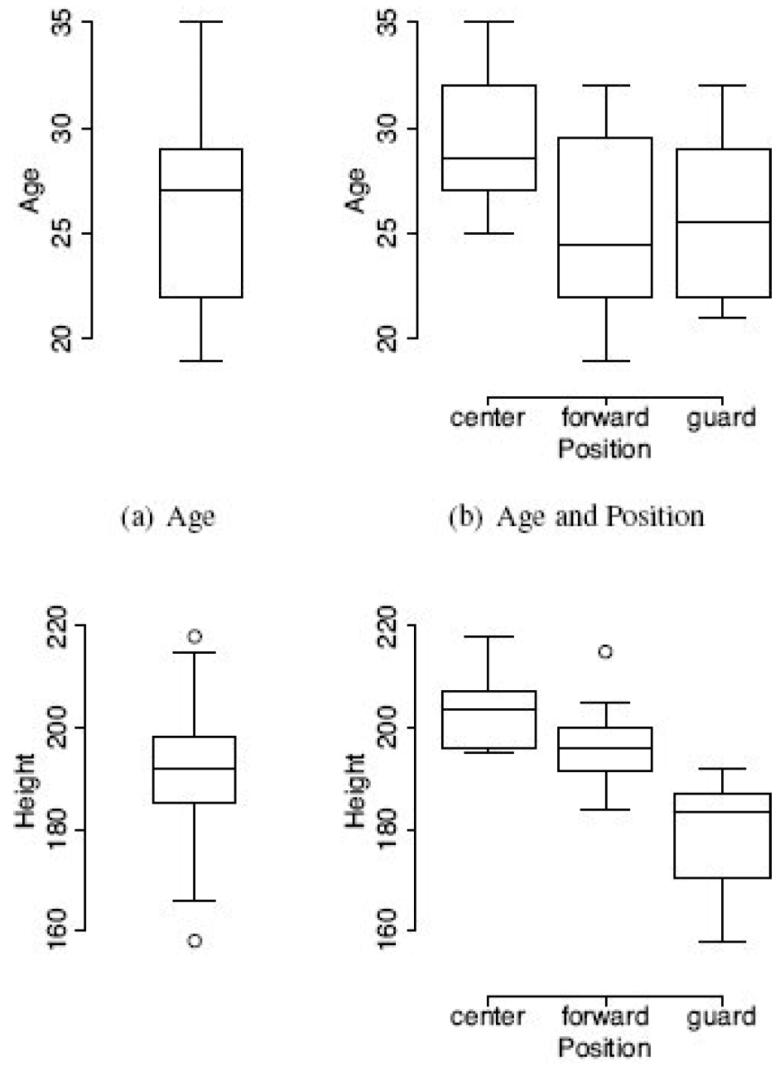
(a) Age (b) Height

(c) Age and Position

(d) Height and Position

Analyzing the Relationship Between Two Variables

- Visualizing a Categorical Feature and a Continuous Feature
  - **Collection box plots**



(a) Age

(b) Age and Position

(c) Height

(d) Height and Position

Issues of Data Quality

- Quality issues due to **invalid data**
  - Caused by errors in the process to generate the features.
  - Fix: correct or regenerate them.

- Quality issues due to **valid data**
  - Exist because of domain-specific reasons.
  - Fix: correct in some cases or do not correct unless required by the trained models, e.g., models cannot be training with missing values or outliers.

Issues of Data Quality

- **Measurement & Data Collection Issues w.r.t. Quality**
  - **Precision:** the closeness of measurements to one another, represented by the standard deviation of the measurements, e.g. repeated measure of body temperature
  - **Bias:** a systematic variation of measurements from the intended quantity measurement, only known when external reference available, e.g. bias in weight measure instrument
  - **Noise:** modification of original values, e.g. distortion of a person's voice when talking on a poor phone, salary="-10".
  - **Outliers:** considerably different from most values in the dataset or unusual with respect to the typical values.
  - **Irregular values:** feature values do not match what we expect, e.g., features with the same value for every instance, (0, 1, m, f, M, and F) to for Male/Female.
  - **Missing values** (Null values): Not measured or Not available, e.g. people decline to give their age and weight, and annual income is not applicable to children.

Issues of Data Quality

- **Main Quality Indicators**
  - **Accuracy:** data recorded with sufficient precision and little bias
  - **Correctness:** data recorded without error and spurious objects
  - **Completeness:** any parts of data records missing
  - **Consistency:** compliance with established rules and constraints
  - **Redundancy:** unnecessary duplicates

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources, e.g., e.g., one rating "1,2,3", another rating "A, B, C"
- Duplicate records

Issues of Data Quality

- **Why Quality is Important?**
  - No quality data, no quality results; “Garbage in, garbage out!”
  - Total data quality control requires a cultural change
  - For most ML projects, *tackling the quality issue at the data source* cannot be always expected; **workaround?**
    - By cleaning the data as much as possible
    - By developing and using more tolerate ML solutions
  - Data quality is relevant to the intended purpose of the ML project, e.g. Does spelling errors in student names really matter when the increase/decrease of student numbers in subject areas over the years are of interest only?

Handling Data Quality Issues

- **Missing Values:**
  - Remove features that are missing in excess of 60% of their values.
  - Replace missing values with an indicator (flag).
  - Impute with mean, median or mode features that < 30% of their values missing.
  - build a ML model that estimates a replacement for a missing value based on the other features.
- **Outliers**
  - Clamp values above an upper threshold and below a lower threshold to these threshold values.