0️⃣

Intro & Visualisation

Status	Done
Files	
Link	
Type	All Around
Course	Data Analysis

These notes are not great but quite difficult to make notes on this topic. They get better for later topics.

General

- Examine data and metadata
- Examine each field/attribute/dimension/column individually
- Examine pairs (correlations)
- Along the way
  - Visualise/summarise
  - Validate assumptions
  - Identify and address anomalies
  - Apply data transformations/corrections
  - Record everything that you are doing

Properties of Data

Structure

- “Shape” of the data
- Easy to manipulate

Granularity

- How fine/coarse is each column
- If aggregated, how?

Scope

- How (in)complete is the data
- Does the data cover the whole area of interest?
- Too expansive? (expansive means covering a large area)
- Right time frame?

Temporality

- Time of data - if data changes with time, when was it collected?
- Meaning of date/time fields? when it was collected vs when it actually happened
- Timezones
- Strange null values? e.g. 1900 or January 1st 1970
- Periodicity?

Faithfulness

- How well the data captures reality
- Is the data to be trusted?
  - Large outliers
  - Negative counts
  - Locations that does not exist
  - Future time date for events in the past
  - Age and birthday dont match
- Entered by hand? expect spelling errors
- Signs of falsification? repeated/fake names/emails

Handling Missing Values

- Ignoring the records (most common) two issues:
  - Not effective when the missing values are spread across many columns
  - Might introduce bias
- Fill manually (tedious)
- Fill automatically
  - Average
  - Average per group (smarter)

- Global constant (”unknown” for example)
- Most probable value (model that predicts the value)

loc vs iloc

loc gets with particular label

iloc gets with particular index

Visualizations for Correlations

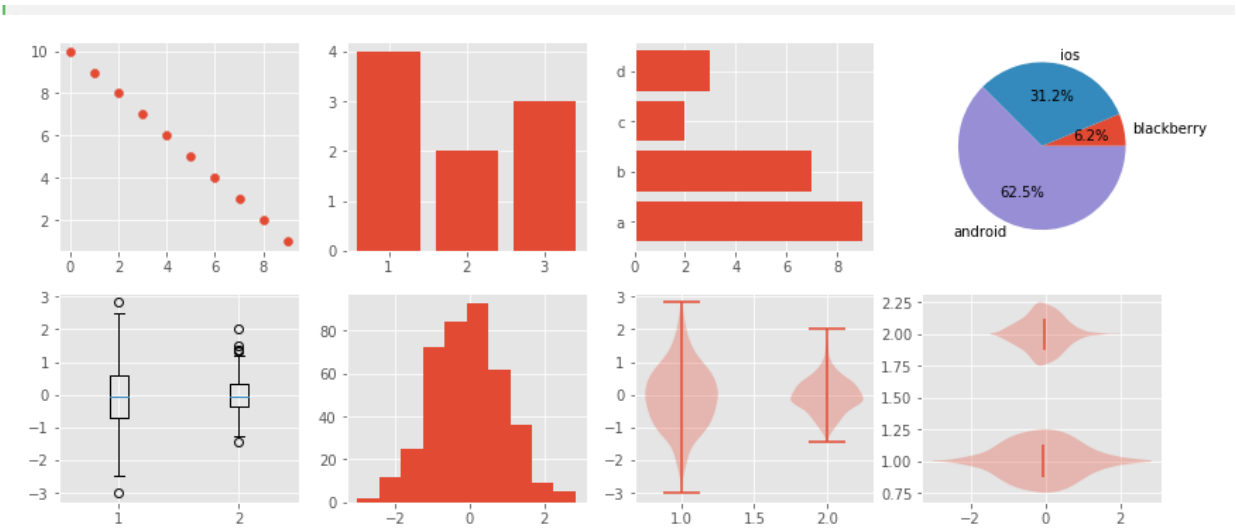
- Scatterplot

Visualisations for raw data, one dimension

- Vertical Bar Chart
- Horizontal Bar Chart
- Pie Chart

Visualisations for distributions

- Boxplot
- Histogram
- Violin Plot



KDE

Relative to a histogram, KDE can produce a plot that is less cluttered and more interpretable, especially when drawing multiple distributions. But it has the potential to introduce distortions if the underlying distribution is bounded or not smooth. Like a histogram, the quality of the representation also depends on the selection of good smoothing parameters.

Variable Types

- Quantitive: numeric value
- Categorical/Qualitative: measures “type” using symbols, names, codes, etc

- Nominal: no order/rank
- Ordinal: natural order

1. Latitude: quantitive, ordinal, continuous

2. Number of siblings: quantitive, nominal, discrete

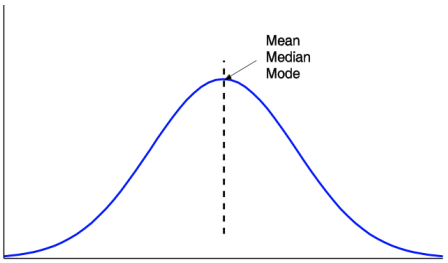
3. Educational level: qualitative, ordinal, discrete

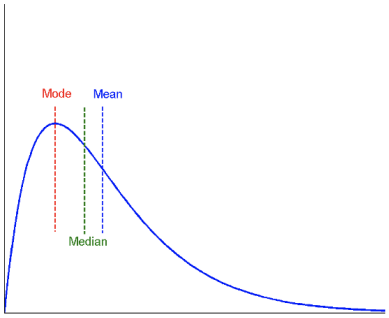
4. College degree: qualitative, nominal, discrete

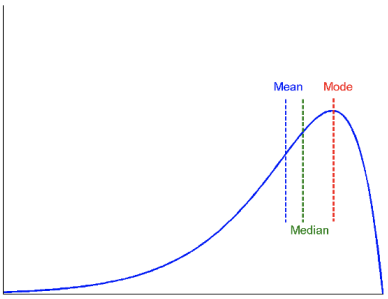
5. Amazon rating for a product: quantitive, ordinal, discrete (individual rating or average rating?)

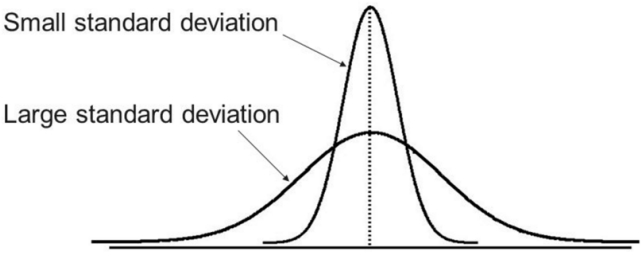
Calculate using numpy

- Mean
- Mode
- Range
- Median
- Percentiles
- Variance
- Std Deviation
- Outlier detection
- Covariance
- Correlation









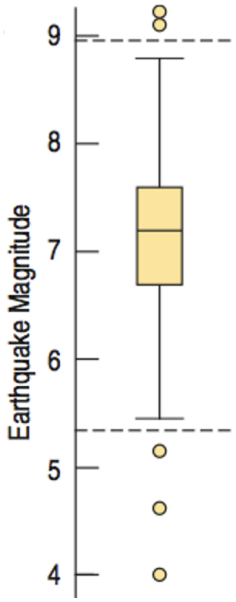
Plot

- Histograms
- Line graphs
- Boxplots

Boxplots

- A Boxplot is a chart that displays the 5-Point Summary and the outliers
- The 5-Point summary provides a numerical description of the data: minimum, Q1, Median, Q3, maximum

Max	9.1
Q3	7.6
Median	7.2
Q1	6.7
Min	4.0



- Frequency tables
- Relative frequency tables
- Bar charts
- Pie charts
- Scatter plot

Correlation

Covariance and Correlation

Covariance and correlation measure of how much two variables change together.

The *covariance* of two variables x and y is given by

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y),$$

where

- μ_x is mean of $x_1, x_2, \ldots, x_n$ and
- μ_y is mean of $y_1, y_2, \ldots, y_n$.

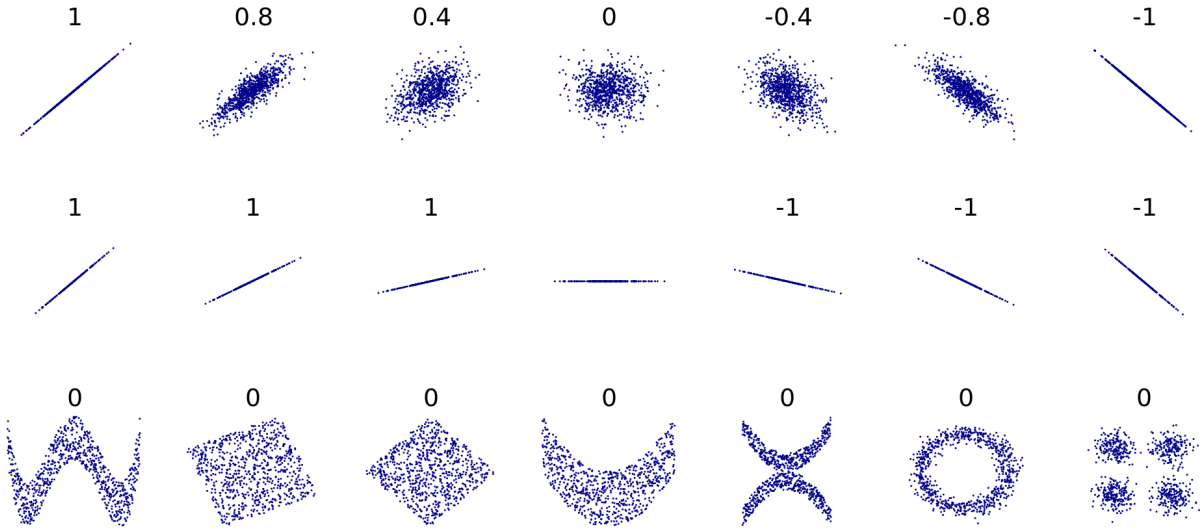
The *correlation coefficient* of two variables x and y is given by

$$corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y},$$

where

- σ_x is std. dev. of $x_1, x_2, \ldots, x_n$ and
- σ_y is std. dev. of $y_1, y_2, \ldots, y_n$.

- Be aware of causation vs correlation, nasty trap

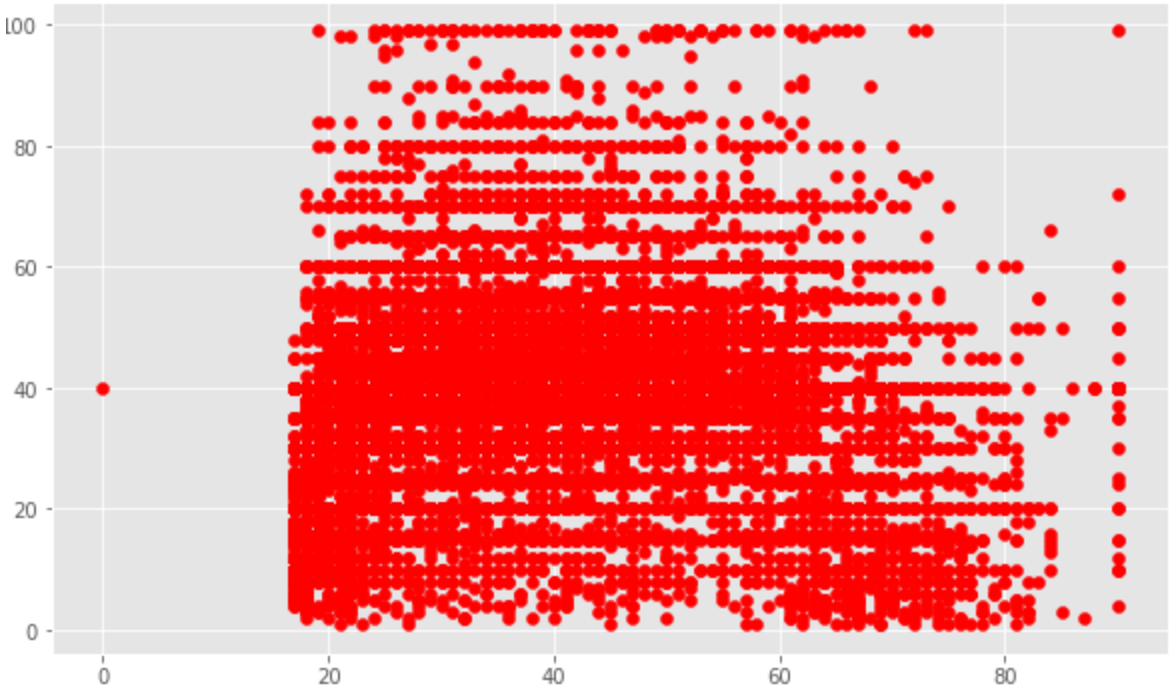



```
* Negative (positive) direction: As one goes up, the other goes down (up)
* No direction means no relationship
* Clear line means a clearly linear relationship
* Dense line means a strong linear relationship
* Extreme values (outliers) can also be found here!
```

```
* To use $corr$, there must be a true underlying linear relationship between the two varia
bles.
* The variables must be quantitative.
* Outliers can strongly affect the correlation. Look at the scatterplot to make sure that
 there are no strong outliers.
* $corr > 0$ --> positive association, $corr < 0$ --> negative association
* $-1 < corr < 1$
* Interchanging x and y does not change the correlation.
* $corr$  has no units.
* Changing the units of x or y does not affect $corr$
```

Overplotting

- Use tiles, or subset of data, or different type of plot, or transparency



Comments

- Avoid 3D graphs

```
* Don’t look for shape, center and spread in a bar chart
* Do a reality check (Choose a histogram bin width appropriate for the datadon’t blindly t
rust calculators or your computer
** A mean student age of 193 years old is nonsense.
** A (human) heart pulse cannot be 800 bpm
* Beware of outliers, the mean and standard deviation are sensitive to outliers.
* Use a histogram to ensure that the mean and standard deviation really do describe the da
ta.
* Don’t compute numerical summaries for a categorical variable (e.g. the mean Social Secur
ity number is meaningless)
* Don't do line plots when not appropriate
```

This is stupid