COMM1190 - Final Notes

Data, Insights and Decisions (University of New South Wales)

COMM1190 - Final Notes

Data Exploration and Visualisation

Data types -

- Cross sectional - observation of many subjects at one point or period of time
- Time series - series of data points in time order
- Panel data - Multiple subjects observed over time - basically combination of cross sectional and times series data
- Textual data - Data based on human language text - e.g. social media posts, customer feedback, legal documents
- Image data - Data consisting of digital images - e.g. photographs, medical images, surveillance footage
- Others

Data quality - six dimensions -

Completeness - data is comprehensive and meets expectations

- I.e. Is all required information available?

Consistency - data across all systems/sourced from different places should reflect the same information

- Are data values the same across datasets?

Conformity - data is following the set of standard data definitions such as data type, size, format

- E.g. date of birth of customer is in the format 'mm/dd/yyyy'

Accuracy - data correctly reflects the real world object or an event being described

- Do data objects accurately represent the 'real work' values they are expected to model? Are there incorrect spellings of product or person names, addresses, and even untimely or not current data?

Integrity - all data in a database can be traced and connected to other data

- Is the data compromised? Is there any missing data?

Timeliness - information is available when it is expected and needed

- E.g. Companies are required to public quarterly results within a given timeframe; companies providing up-to-date information to customers

Summary statistics -

- Mean - location measure
- Variance - dispersion measure
- Standard deviation - measure of risk, can be used to find confidence interval
- Minimum and maximum

- Range
- Covariance
- Correlation

Correlation coefficient -

- A correlation of -1 indicates a perfect negative linear relationship
- A correlation of 1 indicates a perfect positive linear relationship
- A correlation of 0 implies no linear relationship
- The larger the correlation in absolute value, the stronger the (positive/negative) linear relationship

Dealing with outliers -

- Drop the outlier record to avoid severe skewness
- Winsorization - put a cap on data and limit extreme values

Skewness - measure of asymmetry
Kurtosis - measure of peakedness - can tell us whether the data is clustered around the mean

Normally distributed data -

- Bell shaped
- Symmetrical (skewness = 0)
- Kurtosis = 3

Predictive Analysis

Estimating f(X)

- Predict outcome of Y given X
- Explanation/inference - Understanding how Y is affected by X

Regression

- Assumption -> $Y = f(X) + E$
- Y is the outcome, response, target variable
- X = X1, X2,..., Xp are the features, inputs, predictors
- E is the error term capturing the measurement error and other discrepancies

The objective - find an appropriate f for the problem at hand

How to estimate f(X)

- Parametric - make an assumption about the distribution of the data (the shape of f)

- Works fine with limited data, provided that assumptions are reasonable
- Non-parametric - make no assumption about the shape of *f*
  - Needs a large number of observations
  - Less interpretable

Simple linear regression -

- Predict a quantitative response Y based on a single predictor factor variable X
- Approximately a linear relationship between X and Y

$$Y = B_0 + B_1X + E$$

- Use (training data) to produce estimates of B_0 and B_1
- Make predictions given X = x and using the estimates

$$y = B_0 + B_1x$$

Relationship between X and Y - Can be evaluated by Hypothesis testing

Null hypothesis => $H_0 : B_1 = 0$ (There is no relationship between X and Y)

$B_1 =/= 0$ (There is some relationship between X and Y)

- Determine whether B_1 is sufficiently far from 0 using t-statistics

$$t = B_1 / SE(B_1)$$

SE is standard error

- p-value: Assuming $B_1 = 0$, what is the probability of seeing any value equal to t or larger?

  If p-value < 0.05, then reject the null hypothesis

Multiple Linear Regression

- Extend linear regression to accommodate multiple predictors

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p + E$$

- *B_j: Average effect on Y of a unit increase in X_j, holding all other predictors fixed*

Usefulness of the predictors:

- Can be assessed via hypothesis testing -> using F-statistic
- The larger the value of F, the more evidence there is to reject the null hypothesis
- p-value - Given $B_1 = B_2 = \ldots B_p = 0$, what is the probability of seeing any value larger than *F*
- p-value , 0.05 => reject the null hypothesis

Regression vs Classification

Regression - Y is quantitative, continuous

- Examples - sales prediction, house price prediction, stock price modelling, model employment satisfaction

Classification - Y is qualitative, discrete

- Examples - fraud detection, facial recognition, whether someone will default on a debt or not.

Classification trees -

- Set of rules based on input variables to make predictions summarised in a tree
- Simple interpretation/ closely resembles human decision making
- Useful for communicating predictions and results to non-technical stakeholders

Model accuracy can be predicted by -

Accuracy rate = Correct predictions / Number of Points

Logistic regression -

Output is binary: $Y = \{1, 0\}$

<u>Research Design and Experiments</u>

Organisations require answers to questions and input into decision-making

- Research design addresses and maps out constituent parts of analysis
  - Subject matter theory
  - Appropriate data
  - Modelling approach that is appropriate for the data to deliver answers
- Organisations need to answer what-if and evaluation type questions (prescriptive analysis)
  - Experiments with Random Control Trials (RCTs) obtains causal effects
  - Involves causal questions requiring estimates of causal effects
  - What if a change is made? How will that affect future outcomes?
  - What impact did intervention have? Was an implemented policy change effective?
- Design -> Prescriptive analysis
  - Research design: can causal questions be answered by available data and planned modelling approach

Other examples - Obtain data through experiments:

a) Will it be profitable if on-line advertising is increased?
   - Online A/B experiment (split randomly into 2 groups)
b) Will a back-to-work intervention help people get a job?
   - Require a field experiment
c) How much should homeowners living near a chemical plant be compensated for the chemical spill?

- Note: some questions may not be amenable to experiments - too harmful, unethical, unfeasible

- Obtain observational data
- However, there are design issues involved with natural experiments

Randomisation is important to control for confounding factors

Regression as an analytical tool -

- Useful in capturing bivariate relationships
- However, regression cannot be used to answer causality questions or 'what if' counterfactuals.
- Why can't regression be used?: (1) Confounding variables - can lead to omitted variable bias - other factors could be affecting the output; (2) Reverse causality - e.g. what if markets with low sales lead to an increase in advertising (rather than advertising leading to increased sales)?

Causality and notion of 'ceteris paribus':

- Causal effect of x on y: how does variable y change if x is changed but all other relevant factors (u) are held constant

$$Sales_t = a_0 + a_1 sales_{t-1} + u_t$$

- In evaluating an intervention or policy change, think of counterfactual outcomes and what-if questions - e.g. Sales with or without the increase in advertising
- Requires (at a minimum) *x* and *u* to be unrelated

Experiment 1:

Impact of a back-to-work program on employment -

- "If a person is chosen from the population of those looking for work and given access to a back-to-work program, will that increase their chance of employment?"
- Implicit assumption: all other factors that influence employment (experience, ability, local employment prospects, etc) are held fixed

Experiment -

- Choose a group of workers looking for work
- Randomly assign them to access the program (treatment group) or not (control group)
- Compare employment outcomes in next period
- Experiment works because characteristics of people are unrelated to whether they receive the program or not

Experiment 2:

A/B testing of a website landing page

- "If a business rearranges their current website, how much will this change the conversion rate (new customers)?"

- Implicit assumption: all other factors that influence who visits the website are held fixed

Experiment -

- Design a new webpage
- Randomly assign different users to old (A) and new (B) websites
- Compare conversion rates i.e. new customers
- Experiment works because characteristics of users are unrelated to which website is seen
- This kind of experiment is relatively easy to conduct in online environments

Experiment 3:

Policy question -> Importance of investing in education

- "If a person is chosen from the population and given an extra year of education, by how much will his or her wage increase?"
- Implicit assumption: all other factors that influence wages (e.g. experience, family background, intelligence) are held fixed

Experiment -

- Choose a group of people
- Randomly assign different amounts of education to them
- Compare wage outcomes
- *Note* - Random assignment is unfeasible in this case - Experiments are not always feasible or ethical

Conducting RCTs

- Decide on form of intervention (new program vs status quo)
- Determine outcome of interest (employment/conversion rates)
- Decide on randomisation unit (workers/customers/students)
- Determine sample size and randomly assign units to treatment and control groups
- Compare (average) outcomes to determine treatment effects
- Any difference in outcomes can reasonably be attributable to the treatment as other aspects of data controlled by researcher
- Decide on whether to adapt (implement program) or not on basis of findings

Experimental evidence is input into decision-making -

- Even if the RCT yields a significant treatment effect this may not be enough to justify implementation of the intervention
- Does the intervention represent value for money?
- Interventions can be costly so size of any benefit must be weighed against costs

- Null results can be useful -

- A RCT that does not provide evidence of a treatment effect could avoid unnecessary costs
- Informs us to use a different design and not invest in the treatment

- Once implemented, evidence should continue to be collected and interventions refined where appropriate -
  - Interventions may work in one population but not another
  - Replication of core findings across several experiments represents more compelling evidence than findings in only one study.

Case study - SAS valuable career skill

- What if workers choose to participate in the program?
  - If workers based their choice on likely benefits from the program, then they are likely to provide an inflated (biassed) estimate of the treatment effect
- This is selection bias induced by an *endogenous treatment* (*internal treatment*)
- Random assignments can avoid this selection problem

Case study: Causal effect of women as policy makers

- Hypothesis - that women leaders will support policies that women voters care more about
- 'Reserved' indicates reserved for a female village leader
- 'Female' indicates a female village leader
- Irrigation and water are number of new or repaired facilities of this type - men tended to be more concerned about irrigation and women more concerned about water

Gold standard analysis - RCTs often thought of as the gold standard in estimating causal effects

- However, data can misbehave
- Randomisation is effective, but the relevant population may be restrictive
- E.g. Are conclusions appropriate for different populations?
- An RCT may have good internal validity (technical aspect) but lack external validity
- Internal validity - the extent to which the observed results represents the truth in the population being studied
- External validity - the extent to which the findings of the study can be generalised to other situations, people, settings and measures - can it be applied to a broader context?

It is possible that the sample treatment and control groups may differ by chance -

- Good practice suggests adding pre-treatment controls

Challenge of causal inference -

- Sometimes, direct experimentation is not possible/ not ethical (e.g. oil spill)
- Role for non-experimental or observational data
  - These data comes with threats
  - Recall confoundment and selection into treatment

Legal AB Testing -

- Proposed AB testing can be harmful or dangerous to the subjects of a test
- E.g. Trial of a new drug - some subjects receive a placebo. If the drug is life saving, then subjects who receive placebo are more likely to die.

AB testing can also cause reputational risk - this must be understood by those conducting the testing -

- E.g. Facebook testing users on happy and sad content -> legal, but negatively impacted emotionally disadvantaged -> reputational harm for Facebook

Poor design of AB test -

- Consumer law issues - E.g. subjects in group B do not receive a service fit for purpose - could result in legal action
- Ensure it does not breach anti-discrimination laws - E.g. experiments grouping by gender or race if the service received was significantly different to one group

AB testing must account for disadvantages of any subjects and consider whether the alternatives are both acceptable to all subjects of the experiment (this could still lead to reputational risk)

Chemical Spill Project -

A firm has admitted responsibility for a chemical spill

- There is a court case determining damages for residents living near to the chemical spill

Conceptually, this is a causal problem that needs to be solved:

- How much damage did residents incur due to the spill?
- Consider the impact on housing prices, what is the difference in prices now and what would it have been without the spill?
- How can you estimate the spill's impact on house prices?

A possible framework for the analysis is the *Lancastrian view of consumer demand* -

- Products viewed as bundles of characteristics or attributes
- Each of these 'attributes' has an implicit shadow price
- Hedonic regression allows estimates of attribute valuations -
  - Houses are valued for their characteristics - number of bedrooms, locations, etc.

Research designs A -

- Get expert advice on affected area ($near_1$ = 1 if house *i* is in area affected and 0 otherwise)

- Collect data after spill for houses both affected and not
- Compare average price and attribute differences to the spill

$$\log(price_i) = B_0 + B_1 + u_i$$

- I.e. comparing houses near spill vs not near spill
- Doesn't account for confoundment - prices would have been different anyways due to difference in location

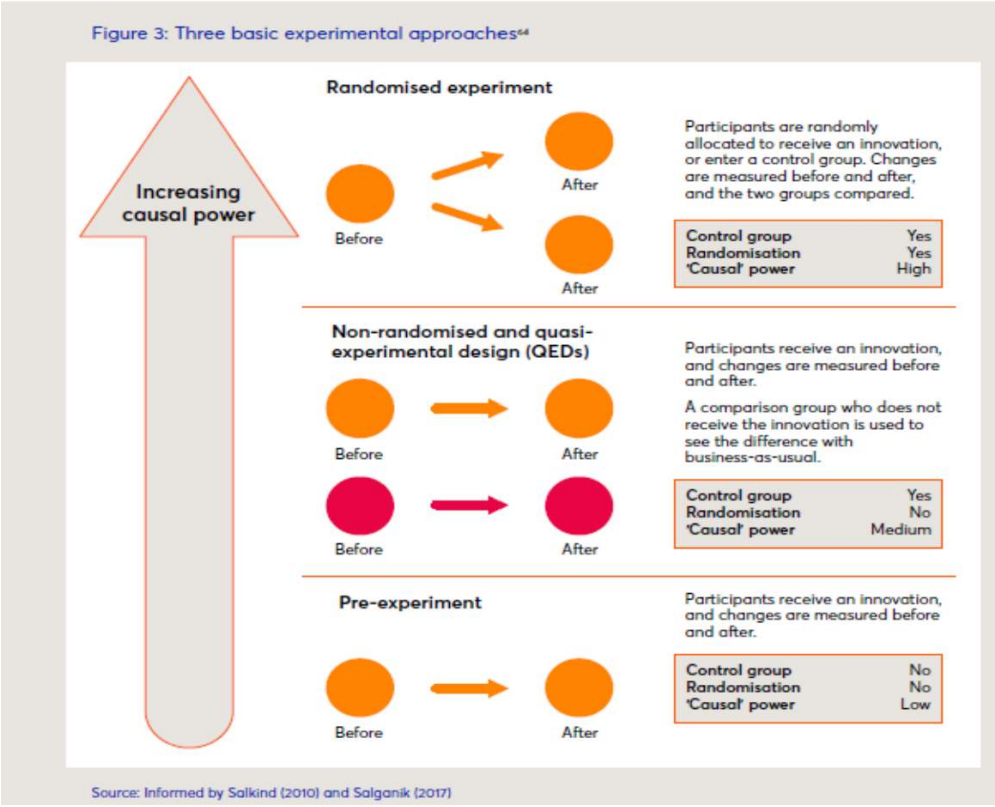
Research Design B -

- Collect data before and after spill for houses in affected area
- Compare prices before and after spill attributing differences to the spill
- Define $after_i$ = 1 if house *i* sold after spill and 0 otherwise
- Parameter of interest would be a_1 in the following model"

$$\log(price_i) = a_0 + a_1 after_i + u_i$$

Research Design C -

- Collect data before and after spill for houses both affected and not
- Compare prices before spill for houses in affected area and those not -> A_b
- Repeat for after period -> A_a
- Now calculate difference of these differences (A_a - A_b) & attribute this to the spill
- Difference in difference (DiD) estimator



Data Ethics

Ethics - 'Moral principles that govern a person's behaviour or the conducting of an activity'
Data ethics -

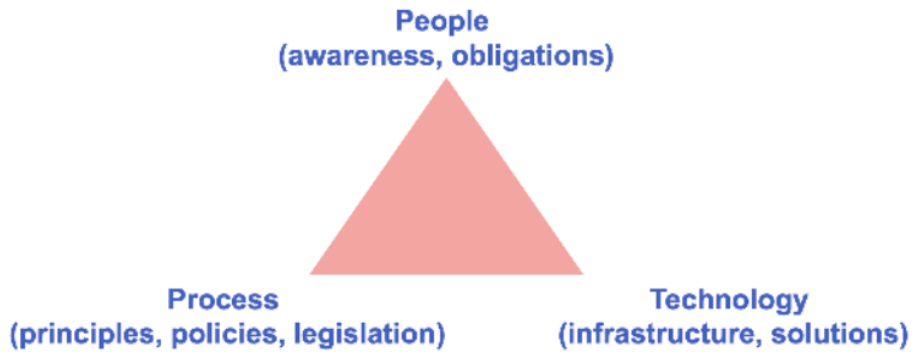
- Moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals
- A new branch of ethics that studies and evaluates moral problems related to data, algorithms, and corresponding practices

Deontological vs Utilitarian -
Deontological - Focuses on moral duties, irrespective of consequences; duty for duty's sake, rules-based approach
Utilitarian - Focuses on maximising positive outcomes for the most people.

Data ethics - Becoming a new source of competitive advantage

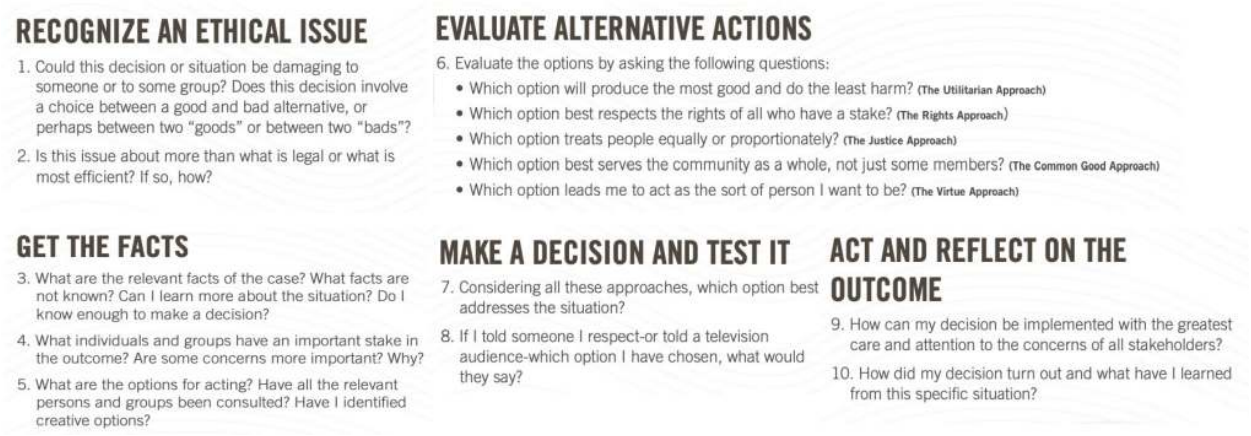
- Responsible business practices - using data for good
- Maintain trust between companies and customers and business partners
- Comply with government and industry regulations
- Enhance business reputation

Data ethics in organisations -



Ethical decision-making framework -

A framework for ethical decisions



Data privacy principles -

- Autonomy - The claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others
- Notice - Inform users about privacy policy, privacy protection procedures
- Choice and consent - Consent from individuals about the collection, use, disclosure, and retention of their information
- Use and retention - Data should be retained and protected according to law or business practices required - e.g. the length of data retention; avoid secondary use of data for other purposes
- Access - Provide individuals with access to review, update, and modify the data about their personal information
- Protection - Data is used only for the purpose stated; de-identifiable of sensitive information; users have the right to opt out for the use of their data
- Enforcement and redress - Provide channels for individuals to report, provide feedback or complain

Data types under protection -

- Identify data - names, address, personal number
- Demographic data - gender, age, education, religion, marital status
- Analysis data - data attributes for which analysis is conducted such as diseases, habits

Dimensions of data protection -

- Anonymity - Users can use resources or services without disclosing their identity

- Pseudonymity - Users acting under a pseudonym may use a resource or service without disclosing their identity
- Unobservability - Users may use a resource or service without others being able to observe that the resource or service is being used
- Unlinkability - Sender and recipient cannot be identified as communicating with each other

Data bias - Types and Mitigation Strategies -

Confirmation bias - The performance of data analysis to prove predetermined assumptions

- Can be avoided by: recording your beliefs and assumptions before starting an analysis; resist the temptation to generate hypotheses or gather additional information to confirm your beliefs.

Outlier bias - Uncomfortable truths are hidden behind a good-enough average

- Can be avoided by: examine the distribution of the sample; use median instead of average; identify and analyse outliers.

Selection bias - Sample is not representative of the population

- Can be avoided by: randomisation; ensuring sampling techniques are appropriate

Survivorship bias - Focus on one side of the story - e.g. only focusing on positives

- Can be avoided by: Develop thorough understanding of phenomenon before data collection

Historical bias - Socio-cultural prejudices and beliefs being mirrored in the analytics process

- Can be avoided by: identifying biases in historical sources; develop inclusive data governance frameworks

Data transparency - Enabling the public to gain information about the operations and structure of the data

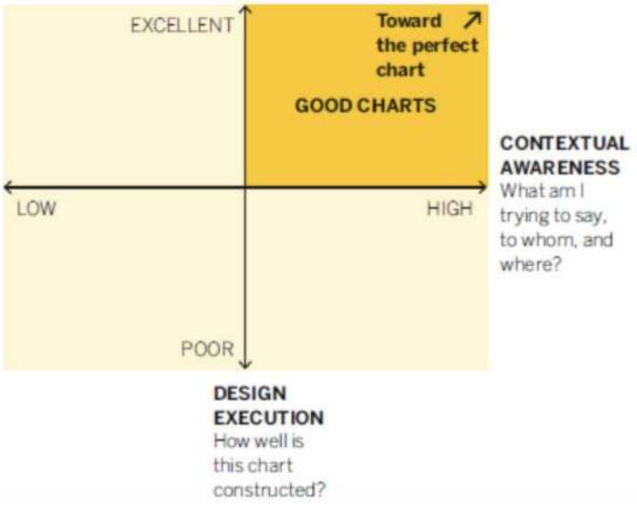
- Understanding how data was selected, recorded, analysed and used
- Being able to access, update, and modify the information

Data Communication

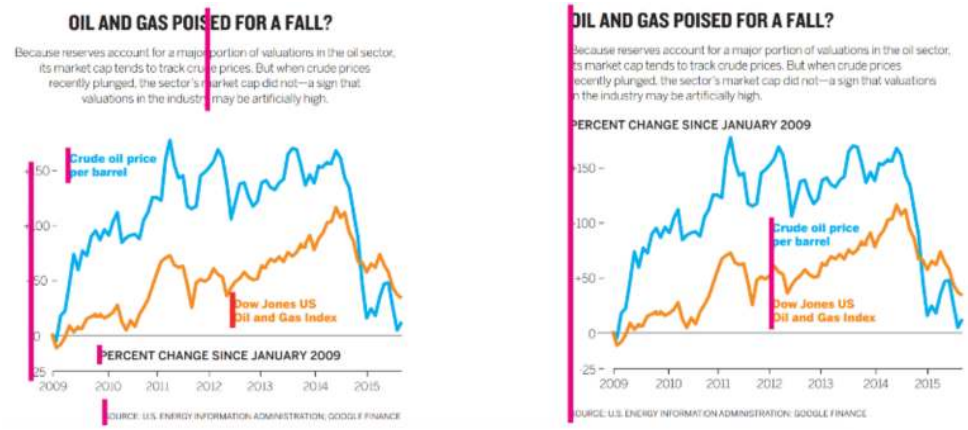
Rules for good charts -

- Don’t use pie charts
- Don’t use geomap plots unless geography is relevant
- Line charts work best for trends
- Do not focus on whether a chart is ‘right’ or ‘wrong’ but focus on whether the chart is good.

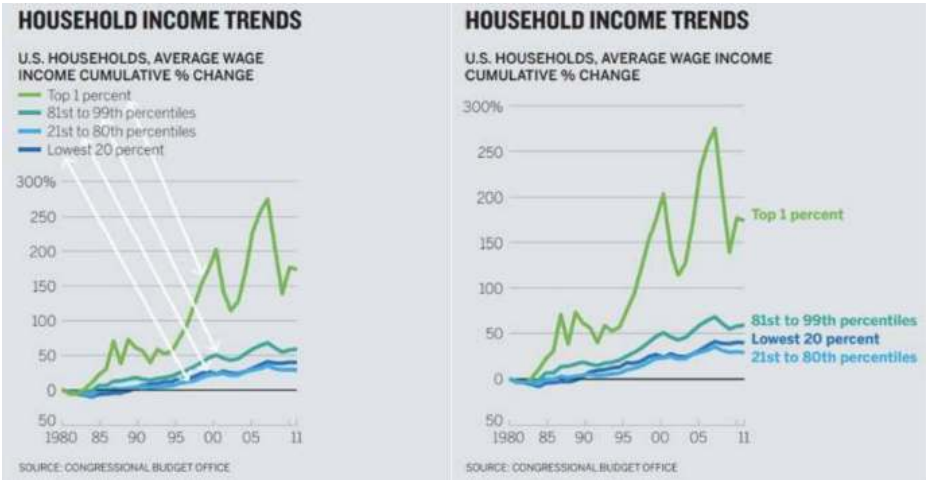
Good charts matrix -



Charts should keep elements aligned -



Good charts limit eye travel -



When is minimalism valuable?

Context: Prototype
Use: Research, individual, informal
Media: Personal screen, paper

Context: "Let's talk about gold prices"
Use: Analysis, informal or formal, one-on-one, small group
Media: Paper, personal screen, public screen

Context: "Gold prices are dropping this year"
Use: Presentation, formal, small or large group
Media: Paper, small screen or large screen

Remove redundancy within key elements -

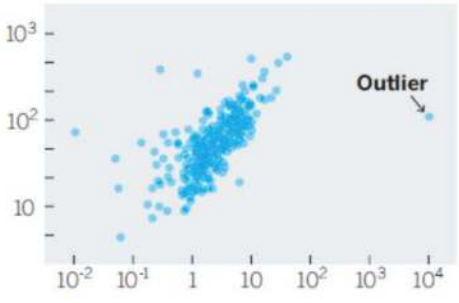
How does a chart hit your eyes?

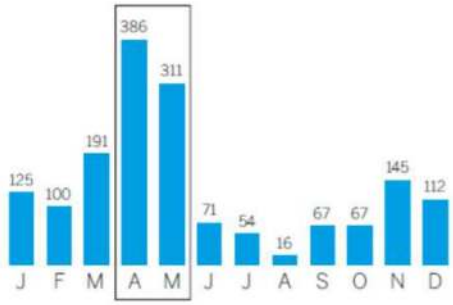
- Ordered colour schemes
- We do not see in order (i.e. left to right) -> we first see *what stands out*

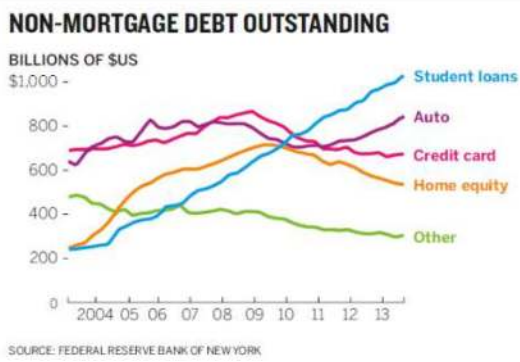
Good charts make a case -

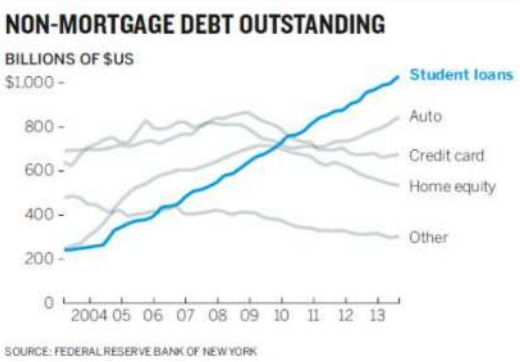
- Competing - attention, resources, financing
- Persuading - pitching clients, swaying opinions, recruiting customers
- Lead to actions

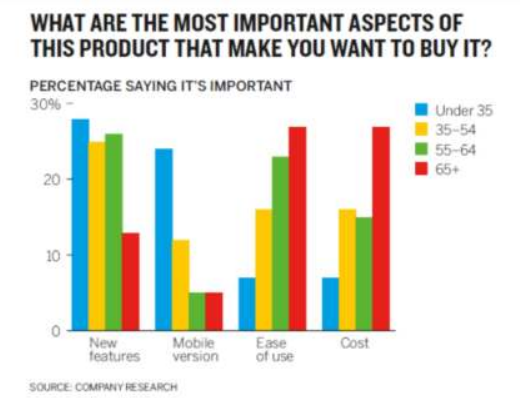
Make your point stand out by *emphasising, isolating, removing or adding info* (in order of the graphs below) -

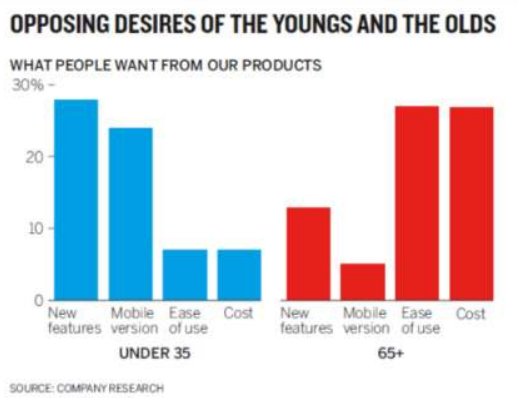






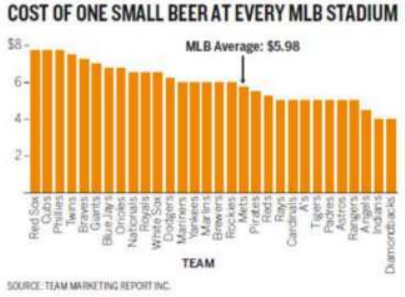






What am I trying to say or show?

I am trying to show the distribution of costs of buying a beer at baseball stadiums.



I need to convince them that . . .

I need to convince them that beer is unreasonably expensive at every single baseball stadium.