Visual Analytics

Ryan G

March 18, 2020

Contents

HouseKeeping

Choosing Between Org and Markdown

This unit won't necessarily be all inside ***R*** so I think the greater versatility of `org-mode` will keep me more organised and allow me to be more versatile.

I will use `org-mode` for this unit as a form of evaluation but `md` for the others (because it's easier to `spin=/=knit` the others directly into `md` without doing a:

```
xclip -o -selection clipboard | pandoc -f markdown -t org | xclip
  ↪ -selectoin clipboard
```

Class Material

The Material for Class is located here:

- file:///home/ryan/Dropbox/Studies/2020Autumn/Visual_Analytics

Textbooks

the textbooks are located in this folder:

- file:///home/ryan/Dropbox/Books/2020Autumn/VisualAnalytics

(Wk 1) Introduction to Data Visualisation

Lecture

Housekeeping

Three will be two report
Basically all you need to do is:

- two reports
- quizz

Shouldn't be too hard to get a HD.

1. Textbook

   It's gonna be a drip fed mess.

2. Assignments

   First one is quite early, like before week 4; 2-people groups.

   Second assignment is on multi-dimensional data

   This was converted from ‘md‘ to ‘org‘ using ‘pandoc -f gfm‘ at time: 2020-03-04T06-00-44

Visualisation

Faster computers means more visuals

1. Definition In a study it was defined as:

   > "The use of computer supported, interactive, visual representations of data to amplify cognition"

   It's all about making data quicker

Nomenclature

- Graphics
  - images using a computer
- Animation
  - graphics objects movement techniques
- Visualisation
  - Exploring, transforming, and viewing data as graphics objects.

Tutorial

- 01. Tutorial Sheet

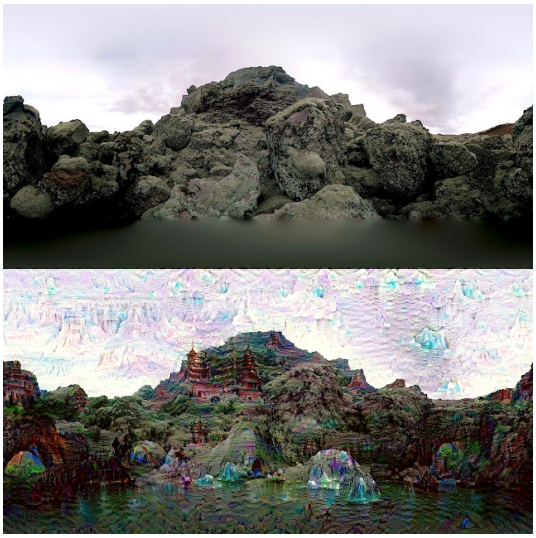
Using *Zathura*

When using *Zathura* the copied text will be placed in the primary * register not the + register, so paste it in with " * p.

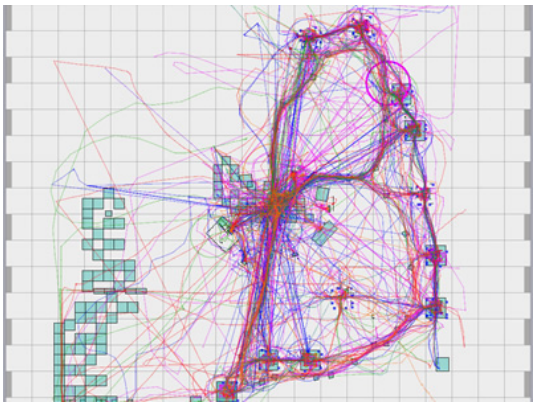
Question 1

1. Problem Visit the website: http://www.visualcomplexity.com/vc/. Have a close look at the available visualisation techniques. In your opinion, which techniques are among the most useful? Or which one is the most pretty visual display? Explain and justify your preference. Note: there is no right or wrong answer in this question.

2. Working This Visual from the *Google AI* blog highlighting the important features detected by a neural network is a unique insight into the *predictive modelling* technique. It is not easy to understand the behaviour of a Neural Network and this visual offers a uniqe insight that could not easily be understood:



This Visual however is arguably one of the most visually striking because of the vibrant colour choice:



Question 2

1. Problem Explore the online demo: http://graphs.gapminder.org/world/. What have you discovered or found from the visualization of the Wealth and Health of Nations data set? E.g. is there any correlation between GDP and Life Expectancy? etc.

2. Working There is a clear positive correlation between income and life expectancy, other features that are readily observable are:

   (a) Regions It can be observed that Europe is a wealthy region (Yellow) while Africa and Asia are poor regions (cyan, red))

   This could be further observed to be a function of distance from the equator, generally regions lose to the equator like Europe, UK, USA are quite wealthy where as regions nearer the equator are more likely to be impovrished.

   Exceptions to this are:

   - Australia due to the rich resources such as coal/uranium and the head start in education owing to the European Descent.
   - Russia is impovrished because the October Revolutions that followed the first

   World War greatly modified there economy relative to the rest of Europe, this decayed into authoritarianism and eventually extreme wealth inequality (with the rise of homelessness and oligarchs) following the collapse of the Soviet Union at the end of the cold-war.

   This could be a functoin of temperature, a higher temperature may lead to :

   - increased spread of bacterial disease
   - increased perishability of food

   (b) Population It isn't possible to say much about the influence population has because the population is measured by region without taking account of size (e.g. China is quite large while Vietnam is relatively small) or habitability (e.g. Japan is very mountainous and hard to farm or ranch on).

   This is made more difficult because the income is measured per person but the population is not the relative space that person might enjoy (i.e. the population density), the GDP per country rather than capita might be more instructive.

   (c) Time Trends

   It appears that Europe (and Japan) have always enjoyed a higher GDP per capita while asia and Africa appear to lag in this respect.

Other than the "reshuffling" caused by the Great Depression, WW I, WW II and the collapse of the soviet union Europe has always enjoyed a higher GDP hence a higher life expectancy.

During the first two world wars (with the exception of Russia and Poland respectively, for obvious reasons) it appears that the influence of currency on life expectancy became even stronger, somewhat unfortunate for other regions given that these were European wars. The great depression also showed a similar effect on this correlation, going from a linear trend to an exponential.

(d) China as an Exception

China appears to not follow a correlation between GDP and Life Expectancy until the new milennia, life expectancy rises following the second world war despite no change in GDP per Capita.

In the 70's China begins to follow this trend moving in a diagonal fashion indicative of a correlation, this is presumably due to manufacturing exported to China perhaps due to the development of the micro-processor in the US.

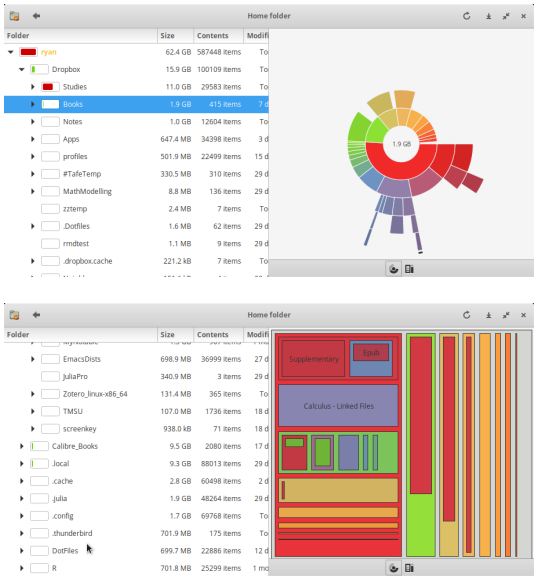
In the 2000's China began to improve GDP per capita more significantly, perhaps due to free market policies, and followed a trend where GDP per capita would strongly correlated with life expectancy.

Question 3

1. Problem Using a search engine, explore 3 applications/projects/tools that use visualisations. You are required to write a short summary about the applications/projects/tools. Why do you think they are significant? What are good and bad about these applications/projects/tools? Etc.

2. Working

   (a) Gnome's Baobab / WinDirStat / KDirStat Gnome' *Baobab*, *KDirStat* and *WinDirStat* are disk usage analysers for Gnome, KDE and *Windows* respectively.



   i. Significance All three tools use a descending list of directories with a bar chart indicating the proportion of disk space consumed by that directory, to the right is a graphic shosing this distribution.

ii. Good Design All three tools have correctly implement an ordered list and bar chart, making it easy to understand and manage disk space on a system.
Gnome's Baobab uses a ring chart with a popup overlay to describe the corresponding directory.
This choice of graph makes it easy to understand which directories are consuming the most space while still having an overview of the structure of the directories, the popup prevents the graph from becoming too busy and showing unnecessary information.
The ring chart will also re-centre following a selection of a directory allowing deeply nested structures to be understood easily.

iii. Bad Design Unfourtunately KDirstat and Windirstat only offer treemap visualisations, this choice of graph is vastly inferior to a ring chart because it can only clearly show a certain amount of information at an overview, it is difficult do understand deeply nested folder structures they won't re-generate the plot without rescanning the drive.

(b) GitHub Visualizer http://ghv.artzub.com/#repo=ranger-assets&climit=100&user=ranger
The Github Visualiser is a way of visualising the proportionate activity of various repositiories of various projects.

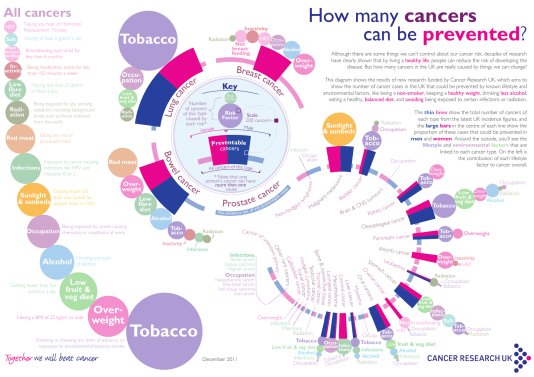
i. Significance This provides a novel way to understand the:
- relative popularity of repos
- The language most composing a repo / project
- How Active a project is generally and over time.

ii. Good Design Using Bubbles to illustrate the is the overall share of a repo in a project provides a quick understanding at a glance.
Using a Time Series Chart over time is an easy way to show trends.

iii. Bad Design The visualisation is far too busy to understand what is going on, for instance the size of the bubbles are not made clear, are they popularity, size, frequency of commits or frequency of clones / pulls of the repo?
For instance the visualisatoin of one of my favourite pieces of software Ranger is such that I can infer nothing from it other than the fact that the web page is written in HTML and the project is written in python, which is totally obvious and not particularly helpful to get a deeper understanding of the *Ranger* project as opposed to say the Midnight Commander Project, although the visualisation for Midnight Commander is significantly better and so this may be a scaling issue.

(c) Cancer Graphic http://web.archive.org/web/20140526084103/http://www.cancerresearchuk.org:80/cancer-info/cancerstats/causes/attributable-risk/visualisation/
This Plot attached here shows the attributes most likely to cause cancer:

```
cd /tmp
#wget "http://web.archive.org/web/20140801035734/http://publications.cancerresearchuk.org/downloads/product/CS_POSTER_ATTRIB.pdf"
command -v pdftoppm >/dev/null 2>&1 || { echo >&2 "command -v foo >/dev/null 2>&1 || { echo >&2 "I require pdftoppm but its not installed.  Aborting."; exit 1; }I require foo but it's not installed.  Aborting."; exit 1; }
pdftoppm CS_POSTER_ATTRIB.pdf  CS_POSTER_ATTRIB -png
mv CS_POSTER*png ~/Notes/Org/Attachments/Statistics/
ls ~/Notes/Org/Attachments/Statistics/CS_POSTER_ATT*png
```



i. Significance This visualisation is significant because it effectively describes both the magnitude and interaction of various risk factors and behaviours on the probability of developing cancer.
   Statistically this can be a very difficult thing to describe and explain but this graphic very clearly shows what to look out or

ii. Good Design Having a key in the centre of the graph makes it very easy to determine what the individual elements of the visualisation mean.
   The relative size of individual risk factors means that at a galnce it is very easy to determine risk factors for cancer.

iii. Bad Design The graph unfourtunately is a little busy but this doesn't appear to be in a way that is disproportionate to the amount of information conveyed.

iv. Inferences An interesting component of this visualisation is that it clearly shows interactions of various items, for example the following are large to moderate risk factors for cancer generaly:

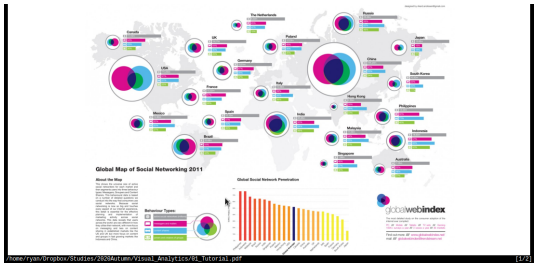
   - Overweight (BMI $>$ 25 kg / m^2, distninct from obesity which is $\approx > 30$)
   - Inactivity

   While HRT is a very small risk factor for all types of cancer.
   This clearly illustrates that HRT may be medically appropriate in men despite the historical belive that such treamtent was correlated with prostate cancer in men [**bell2018**] . Recent studies suggest that this might not be the case [**loeb2017**] and HRT is (somewhat obviously) correlated with increased physical activity, weight loss [**traish2014**] and general health metrics [**saad2020**] .

This visualisation clearly shows that this treatment might generally lower the risk of cancer in men and provides a simple way to convey complex interactions between attributes and the complex statistics involved in this type of research in a way that clearly shows the important facts of the matter..
Moreover this plot also only shows HRT as a risk factor in cancers that tend to overwhelmingly affect women, a distinction that is very important in that area of medical science.

Question 4

1. Problem For the following visualisations, in your opinion, are they good or bad; Justify the answer:
2. Working
   (a) Plot A



This plot is overly busy but also not very descriptive, meaning at a glance it is not possible to determine what the differences between different regions actually are.

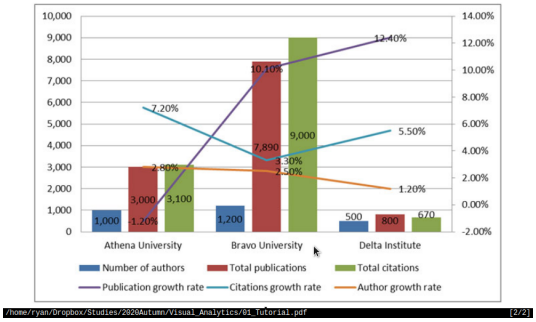
Upon closer inspection it is not clear what the bar charts are trying to illustrate, clearly the venn diagram is trying to illustrate the proportion of communication and overlap, relative to thhe number of users. This is an interesting way to try and show the interactions between the different communication strategies but the choice of a Venn diagram was misguided. A venn diagram can only represent 3 sets with circles [1] and this has seemingly artificially restricted the number of potential behavriou types that the plot has tried to convey to such an extent that they may as well be disregarded entirely because they are just not descriptive enough to understand.

Interestingly Eastern countries use a combination of content/messges while western countries tend to use or the other, this could be related to the pictographical written laguage implemented by Eastern / Asian Countries and could lead to insights on human behaviour, unfourtunately the plot does not clearly describe the meaning of the colours and so no inferences can be easily made.

Moreover the size of the circles are relative to the total number of users, this is somewhat misleading because the plot shows that 33% of the USA uses Social Media while only 15% of China uses social media, countries/regions are not uniformally distributed or constructed so choosing to use regions as a delimiter when using absolute size of users is potentially misleading. This is partially addressed however by the *Global Social Network penetration* bar chart below the plot.

[1] combinatorics - Why can a Venn diagram for $4+$ sets not be constructed using circles? - Mathematics Stack Exchange

(b) Plot B



- The amount of text overlayed on this graph makes it difficult to read.
- The Growth rate of the publication should not be represented as a line because

it indicates some type of continuous connection between the universities, it should be represented as either another column or ideally a seperate time series plot should be produced:

- Such a plot should show the market share and allow the growth rate to be

interpreted from the slope of the line in an organic fashion, that way the different universities could have both market share and growth rate compared between each other without trying to mentally visualise a cumullative summation.

Question 5

Visualisation of the CoronaVirus:

- http://rocs.hu-berlin.de/viz/sgb/

1. Problem Optional: If you still have time, play around this website to see how visualization help to find patterns: http://rocs.hu-berlin.de/viz/sgb/ (Coronavirus Geographic and Network visualisations)
2. Working

   :HideRef:

(Wk 2) Human Visual Perception

Refer to:

- Lecture #2
- Tutorial #2

Lecture ATTACH

Refer to the lecture slides Here and the required readings:

- Lucy Park - Visualisation (Local File)

  - Online Link

- PDF of Human Visual Perception Plot
  - HTML of Perception in Visualisation
  - Online Link

The Tutorial is here

Last Lecture

Good Visualisations are subjective to a degree, but there are basic rules that can be taken from psychology.

Module Outline

- Perception and Cognition
- Human Vision
- Colour
  - Humans can recognise 8 colours at a time so use no more than 10 at a time
    * Ideally use 7 ± 2
- Gestalt Laws
- Visual Encoding

Human Vision

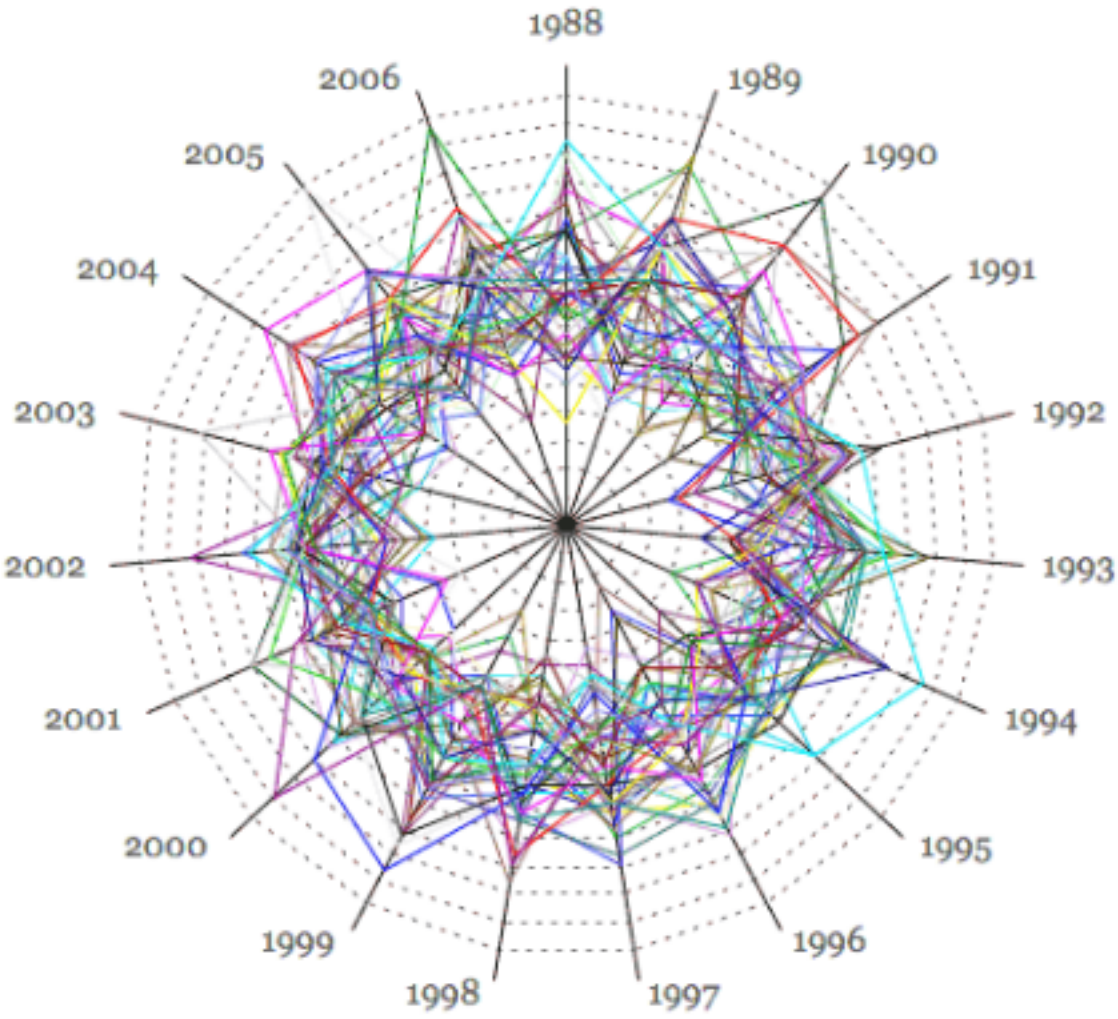
What you see depends when you look at something you're looking at something depends on what you know about. What we see depends on our goals and expectations.

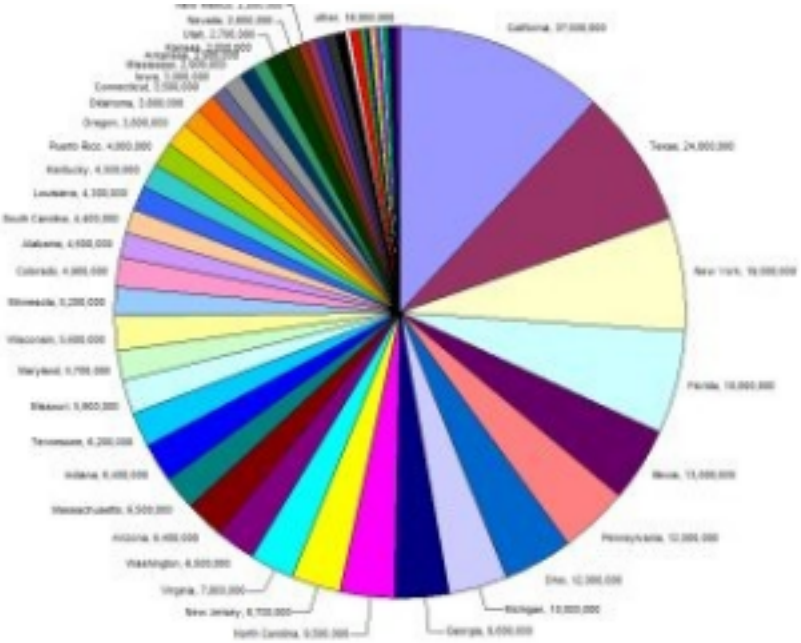
Humans are good at relative value judgements but poor at absolute judgements, whereas computers are the opposite, our goal is to leverage those differences.

Colour Vision

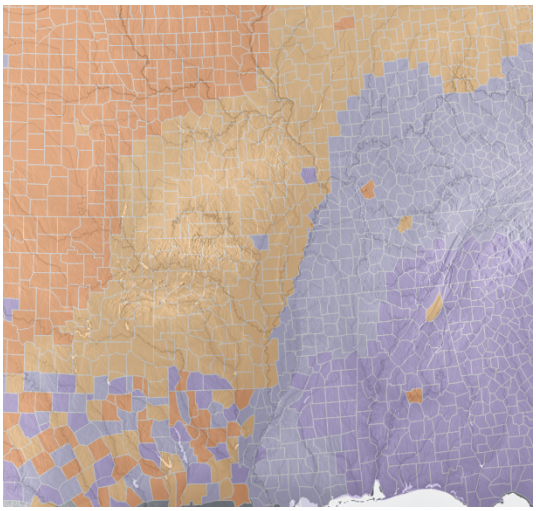
Humans can only identify about 8 colours on a plot, so try to use 7 ± 2 and less than 10, so for example, look at the following:

Lotto numbers, like a star





There's just too much going on to figure out what's going on, whereas look at ColorBrewer:



1. Colour Blindness

Preattentive Processing

Humans analyse images very rapidly, on the order of 200 ms, it's quite accurate and processed 'in parallel' by a low level form of consciousness. This occurs before a person pays attention and is independent of distraction, The opposite of doing a where's waldo.

This can be aided by highlighting things in colour for example if I had a bunch of numbers:

```
sample(1:9, replace = TRUE, size = 30)
```

How many 3's are in there? whereas if one was highlighted:

```
:  [1] *3* 9 8 9 1 1 9 6 2 1 2 2 4 9 7 5 *3* 8 6 5 5 6 3 5 2 1 1 8 7 7
```

They can be rapidly identified.

1. Tasks People can:

   - detect dargets
   - boundary detectoin
   - etc.

   Whereas machine learning algorithm's to do that are 'cutting edge'.

2. Animatoin Be careful with animation, it can be distracting

TODO Visualisation Theory

Read the following article for the tutorial

- Perception in Visualization.

Gestalt Principle

German psycholoogists in the early 1900's attempted to understand perception and formed the *Gestalt School of Psychology* (Gestalt is german for shape/form)

	Categorical		Quantitative	
Scale	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Distinct feature	Distinct categories	Ordered categories	Meaningful distances	Absolute zero
Operations	Equality / Inequality	Smaller than / Larger than	Addition / Subtraction	Multiplication / Division
Example: User study	Participant number	Order of participation	Scale rating (e.g. 1 to 5)	Response time in seconds

Visual Encoding

Effective Encoding of Data Mackinlay 1986 will be in tutorial 2

1. Scales of Measurement

   Basically you just want to read this:

   - Lucy Park Visualisation

   And pay attention to things like this:

2. **TODO** Figure out how to attach websites

Evaluating a Visualisation

In order to evaluate a visualisation, consider the following (taken from here)

- Expressiveness
  - Do the mappings show the facts and only the facts?
    - Are visual mappings consistent? (e.g., respect color mappings)
- Effectiveness
  - Are perceptually effective encodings used?
  - Are the most important data mapped to the most effective visual variables?
- Cognitive Load (Efficiency)
  - Are there extraneous (unmapped) visual elements?
- Data Transformation
  - Are transformations (filter, sort, derive, aggregate) appropriate?
- Guides (Non-Data Elements)
  - Descriptive, consistent: Title, Label, Caption, Source, Annotations
  - Meaningful references: Gridlines, Legend

Potential Assignment resources

This Visualisation is potentially something that I could use for one of the assignments.

Tutorial

The Tutorial is here

All Exercises

Tutorial Exercises:

1. What are the differences between perception and cognition?
2. Why human vision is important for data analysis? What are the pros and cons of using human visual system versus machine or automated analysis?
3. Why colours are important in visualisation?
4. List at least 5 main things that we should be aware when using colours in visualisation.
5. Using a search engine, explore 3 good visualisations that use colours to represent information. Why are the colours used effectively in these visualisations?
6. Test your eye with colours: http://enchroma.com/test/instructions/

7. Why do you think pre-attentive processing is important in visualisation? Can we combine features (e.g. colour, size and shape) to enhance pre-attentive analysis?
8. What are the main aspects of Gestalt Laws?
9. Read the slide Effective Encoding of Data in Lecture Note 2 (https://www.lucypark.kr/courses/2015-ba/visualization1.html), if a dataset has quantitative value, what 5 most important attributes or encoding should we consider? Similarly, if a dataset has nominal value, what 5 most important attributes or encoding should we consider?
10. Read the following article: http://www.csc.ncsu.edu/faculty/healey/PP/index.html

Question 1

- 1. What are the differences between perception and cognition?

That basic distinction between pereption and cognition is that perception is tied to observation and sensation whereas cognition is related to information is related to the understanding and processing of ideas.

1. Perception Perception is identifying and understanding a sensory input, such as perceiving:

   - an angle as obtuse or acute
   - an object to be round or sharp by touch or sight
   - an accent to be British, German, Chinese etc. by sound.
   - an object to be falling off the table.

   Perception is inherently dependent on *experience*, for example a person without exposure may not be able to perceive differences in accent.

   It is unlikely that two different people will perceive things in the exact same way.

   Generally perception is not something done consciously.

2. Cognition Cognition is processing information in an analytical way, working through a problem using prior training, it is done consciously.

   Examples of cognition include:

   - Determing the shape of a plot follows a certain model (e.g. polynomial, exponential/logarithmic, ARIMA, White Noise, etc.)
   - Recognising an object to be a 12 mm nut from brake caliper as opposed to a small metallic object.
   - Deciding that a person probably spent time in particular countries or regions given the languages they speak and accent they have.
   - Determining that an object that is falling can be modelled with calculus.

   Cognition depends on training, the majority of people can learn to work through a problem in a particular way.

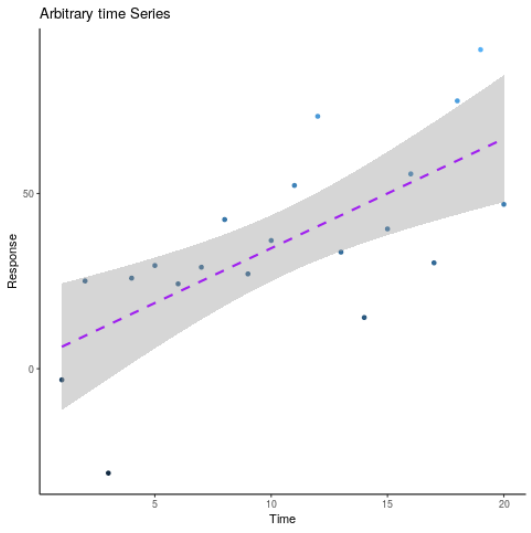
Question 2

- 2. Why human vision is important for data analysis? What are the pros and cons of using human visual system versus machine or automated analysis?

Human vision is important for data analysis because often times there is too much data to understand just by looking at it. More over inferences from that data often cannot be made without visualising or summarising it.

Often it is possible to perceive change by looking at a visualisation but not possible when looking at raw data, for example look at the following plot: than trying to analytically derive a process, for example the following is easy to perceive:

```r
library(tidyverse)
x <- 1:20
y <- x*3+2 + rnorm(length(x), 0, 15)
data <- data.frame(x,y)


ggplot(data, aes(x = x, y = y, col = y)) +
             geom_point() +
             guides(col = FALSE) +
             theme_classic() +
             stat_smooth(method = "lm", lty = 2, col = "purple") +
             labs(title = "Arbitrary time Series", x = "Time", y =
             ↪  "Response")
```



An individual can clearly perceive that there is a linear trend with a rate of about 3 and upon closer inspection the standard error shading shows that the data is roughly $\pm$ 15.

Looking at the raw data however:

1	-3.2
2	25
3	-29.8
4	25.8
5	29.4
6	24.2

It would only be possible to determine that the rate is indeed linear as opposed to polynomial by performing a statistical test and the rate would need be determined Analytically by considering the residual:

$$w_0 = \frac{\sum_{i=1}^{n}[y_i]}{n} + \frac{w_1 \sum_{i=1}^{n}[x_i]}{n} \tag{1}$$

$$w_1 = \frac{\sum_{i=1}^{n}[x_i y_i] - \frac{\sum_{n}^{i=1}[x_i]\sum_{n}^{i=1}[y_i]}{n}}{[\sum_{i=1}^{n}[(x_i)^2] + \frac{(\sum_{i=1}^{n}[x_i])^2}{n}]} \tag{2}$$

Humans are good at relative value judgements but poor at absolute judgements, whereas computers are the opposite, our goal is to leverage those differences to our advantage.

1. Human Visual System and Automated Analysis The pros and cons of Human Perception and Automated Analysis are best highlighted by considering the difficulties of unsupervised learning in the realm of Machine Learning.

   Unsupervised Machine Learning algorithms such as clustering and PCA (in 2 and 3 dimensions obviously) are very easily performed by humans, but implementing such algorithms automatically and analytically can be quite complex and in some cases resource intensive.

   This is very similar to the proximity *Gestalt* law discussed in question 8 below at .9.

   An advantage to the Human visual system is that the analysis can be done rapidly and requires no prior programming and maybe to a degree a human can be influenced quite rapidly in how the data is analysed or perceived whereas programming a machine learning algorithm or even a deductive algorithm can be complex and time consuming.

   The advantage to using Automated analysis and algorithms is that an algorithm:

   - is unbiased to the origins or nature of the data
   - can be extended to disjoint, large data that might have patterns outside what humans would be expected to recognise
   - can be extended to multiple dimensions
     - For example using KNN analysis might involve computing distances in higher dimensions
     - Or interpreting data with many variables, PCA can 'flatten' out variables that don't significantly contribute to the variance of the response.

   (a) Summary

   - Humans can make rapid and accurate relative judgements but these judements are susceptible to bias and influence.
   - Machines can make unbiased consistent judgements based on an algorithm (or atleast a machine learning algorithm) but cannot out of the box make a relative value judgement without human guidance.

Question 3

- 3. Why colours are important in visualisation?

Colours are important because they can be used to illustrate a discrete or continuous dimension of data, this is particularly useful for comparing different relationships as they relate to different populations.

Question 4

- 4. List at least 5 main things that we should be aware when using colours in visualisation.

1. (1) Number of Colours Humans can only identify about 8 colours on a plot, so try to use 7 ± 2 and less than 10.
2. (2) Data Type

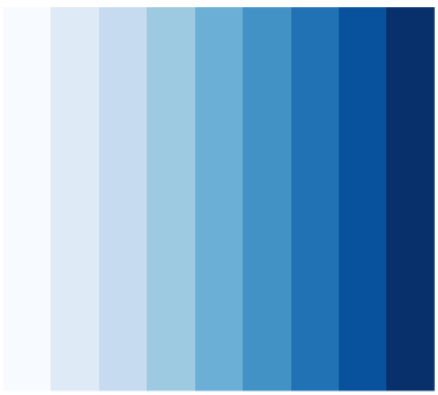
   Data may be ordered and may be continuous or discrete:

	Ordered	Unordered
Continuous	Colours should use a smooth divergent/sequential pallet	NA
Discrete	Colours should use a small number of sequential colours	Colour

   pallet should be a small number of distinct colours |

   If data is continuous (and / or ordinal) then colours that follow a pattern should be used in order to illustrate that pattern, for example:

```
library(RColorBrewer)
library(tidyverse)
# colorRampPalette(brewer.pal(9, "Blues"))(100) %>% plot()
# my_cols <- brewer.pal(7,"Greens")
# par(pty = "s", mai = c(0.1, 0.1, 0.4, 0.1))
# display.brewer.pal(3,"Accent")
 display.brewer.pal(9,"Blues")
```



Blues (sequential)

If data is discrete then discrete and distinct colours should be used, if that data is also non-ordinal then those colours should avoid having a pattern, for example:

```
library(RColorBrewer)
library(tidyverse)
# colorRampPalette(brewer.pal(9, "Blues"))(100) %>% plot()
# my_cols <- brewer.pal(7,"Greens")
# par(pty = "s", mai = c(0.1, 0.1, 0.4, 0.1))
display.brewer.pal(3,"Accent")
# display.brewer.pal(9,"Blues")
```

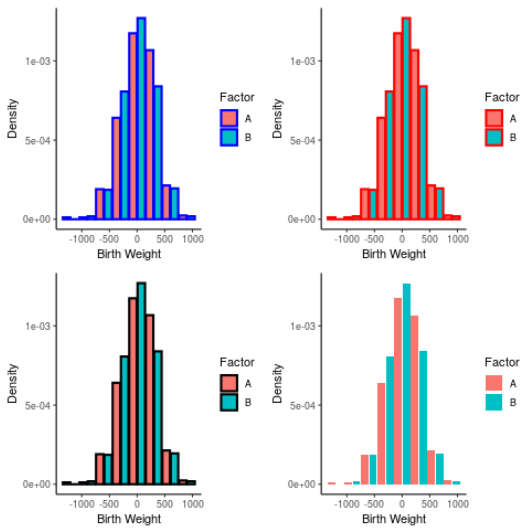
Accent (qualitative)

3. (3) Viewing Medium A visualisation may have different levels of perceived brightness depending on how it's viewed, for example a visualisation made on a pc with a black background (i.e. dark mode) may lead to a plot looking overly saturated, if that plot was printed the distinction between various factors may not be clear in graph.

4. (4) Contrast If a plot is converted to greyscale in order to print, there may be insufficient contrast to perceive differences, this can be avoided if attention is paid to ensure good 'seperation' between elements regardless of contrast.

5. (5) edge enhancement using colours to highlight edges may make a plot harder or easier to read, for example consider the following histograms:

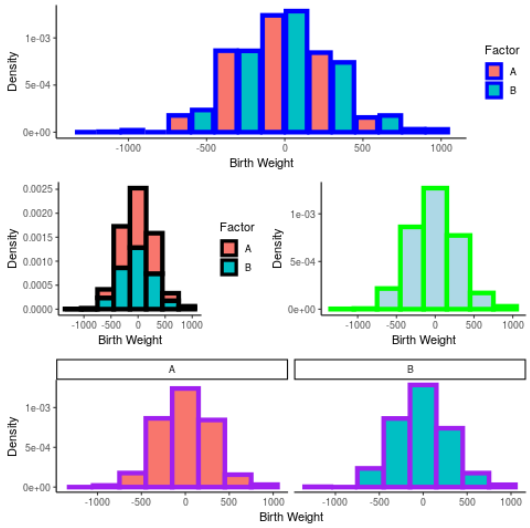
```r
library(tidyverse)
library(gridExtra)
a <- rnorm(1000, 3, 300)
b <- rbinom(length(a), size = 1, prob = 0.3)
birthwt <- data.frame("A" = a, "B" = b )
birthwt_pretty <- birthwt
birthwt_pretty$B <- ifelse(birthwt$B, "A", "B")

hist <- ggplot(birthwt_pretty, aes(x = A, fill = B,  y = ..density..)) +
        theme_classic() +
        labs(x = "Birth Weight", y = "Density") +
        guides(fill = guide_legend("Factor"))

plots <- list()

plots[[1]] <- hist + geom_histogram(position = "dodge2", col = "blue", binwidth = 300, lwd = 1)
plots[[2]] <- hist + geom_histogram(position = "dodge2", col = "red", binwidth = 300, lwd = 1)
plots[[3]] <- hist + geom_histogram(position = "dodge2", col = "black", binwidth = 300, lwd = 1)
plots[[4]] <- hist + geom_histogram(position = "dodge2", binwidth = 300, lwd = 1)


layout <- matrix(c(1:4), byrow = TRUE, nrow = 2)
grid.arrange(grobs = plots, layout_matrix = layout)
```



This effect can be made more pronounced when the plots have different scales an are arranged

relative to each other:

```r
a <- rnorm(1000, 3, 300)
b <- rbinom(length(a), size = 1, prob = 0.3)
birthwt <- data.frame("A" = a, "B" = b )
birthwt_pretty <- birthwt
birthwt_pretty$B <- ifelse(birthwt$B, "A", "B")

hist <- ggplot(birthwt_pretty, aes(x = A, fill = B,  y = ..density..)) +
        theme_classic() +
        labs(x = "Birth Weight", y = "Density") +
        guides(fill = guide_legend("Factor"))

# hist + geom_histogram(position = "dodge2", col = "blue", binwidth = 300)

plots <- list()

# Dodge
plots[[1]] <- hist + geom_histogram(position = "dodge2", col = "blue", binwidth = 300, lwd = 2)

# Overlay
plots[[2]] <-  hist + geom_histogram(binwidth = 300, col = "black", lwd = 2)

# Single Histogram
plots[[3]] <- hist + geom_histogram(binwidth = 300, col = "green", aes(group = 1), fill = "lightblue", lwd = 2)

# Facet Grid
 plots[[4]] <- hist + geom_histogram(binwidth = 300, col = "purple", lwd = 2) +
        facet_grid(. ~ B) +
        guides(fill = FALSE)

layout <- matrix(c(1, 1, 2, 3, 4, 4), byrow = TRUE, nrow = 3)
# arrangeGrob(grobs = plots, layout_matrix = layout)
grid.arrange(grobs = plots, layout_matrix = layout)
```



Question 5

- 5. Using a search engine, explore 3 good visualisations that use colours to represent information. Why are the colours used effectively in these visualisations?

Further Information on correct colour choice in plots can be located:

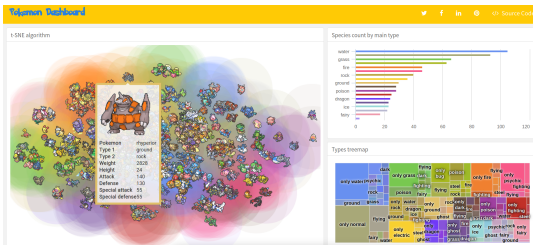
- DataQuest
- GraphiQ
- DataWrapper

And many good examples can be found on the *RStudio Shiny Dashboard* homepage.

1. Pokemon dashboard The Pokemon type dashboard uses colours effectively because it uses discrete and distinct colours to represent discrete variables.

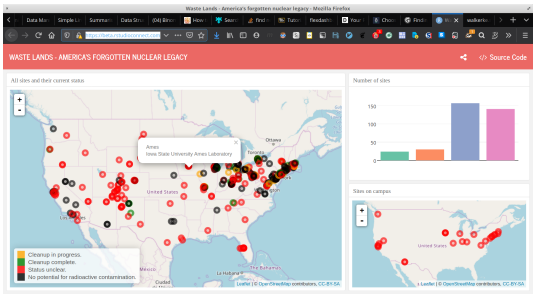
   The colour choices are an intelligent choice of sufficiently distinct colours that are also differently related to the information to make the plot easy to interpret.

   The colours choses are also saturated enough to allow transperancy to be mapped to frequency and for the treemap to be sufficiently distinct.

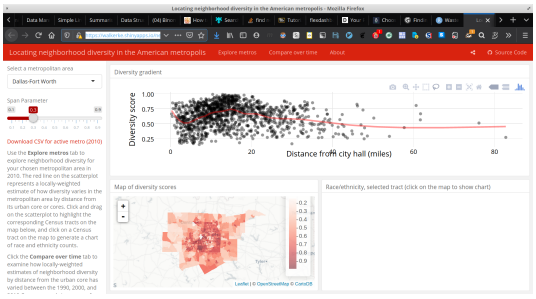


2. Nuclear Waste Sites The Nuclear Waste Dashboard uses colour effectively because it uses a different colour pallette for the two types of plots which represent the number of sites in each category and the geographical location of any such site in that category.

   Tying a colour to a variable is important to visualise data without confusing the intended audience.

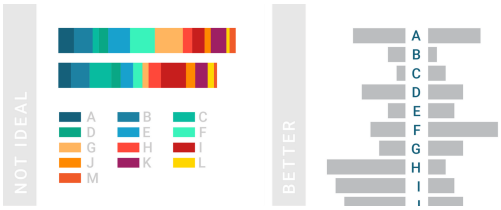


3. Neighbourhood Diversity The Neighbourhood Diversity visualisation uses colour effectively because the colour choses is a continous sequential pallete increasing in saturation, this makes interpreting continuous data easier because the pallete has a one-to-one correspondence with the density of the observation.



4. Using Colours

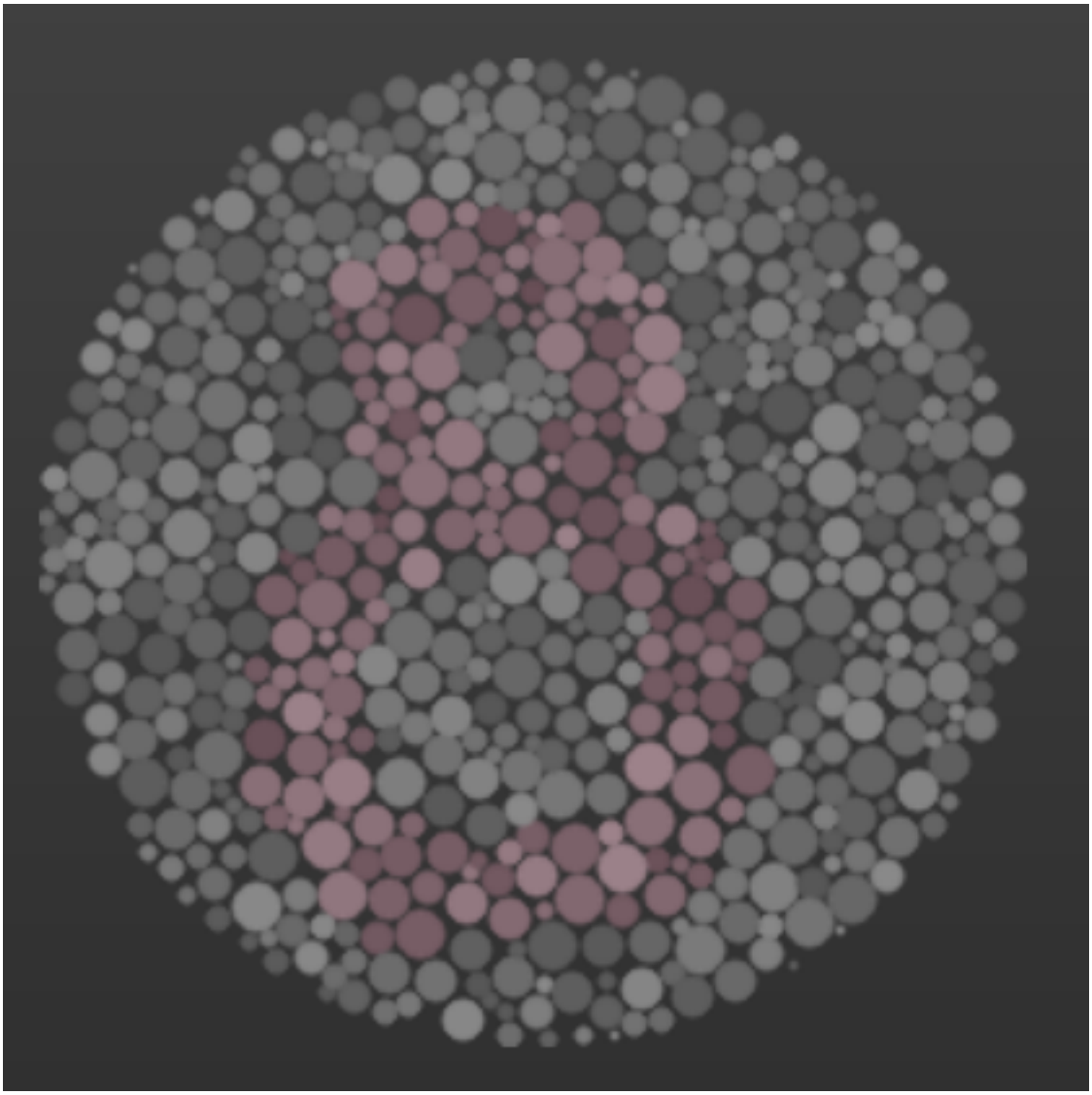
   The following graphic is an exemplar graphic from *DataQuest* but it shows how it is often desirable to restructure a visualisation based on the data as opposed to overusing colours which can be difficult to interpret in high amounts.



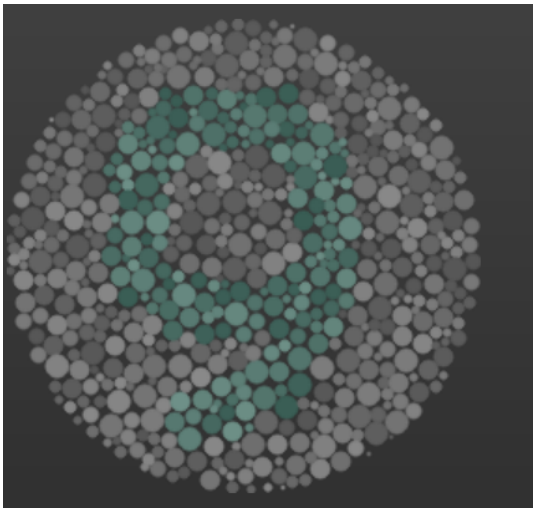
Question 6

- 6. Test your eye with colours: http://enchroma.com/test/instructions/

I already know I'm slightly red-green colourblind, (Mild Protan), the test only confirms it, for instance I don't see any number here:



or here:



Question 7

- 7. Why do you think pre-attentive processing is important in visualisation? Can we combine features (e.g. colour, size and shape) to enhance pre-attentive analysis?

Pre-Attentive processing is important because it allows for a plot to be interpreted immediately.

A visualisation may be rendered less effective or not effective at all if it is unable to convey certain amounts of information rapidly, for instance if the cognitive load to interpret a pattern is too high, sub patterns or more complex patterns may be lost in the noise of the data.

Question 8

- 8. What are the main aspects of Gestalt Laws?

The Gestalt Laws provide that:

1. Grouping
   - Objects close together will be perceived as groups
     - This is an example of humans performing something simple that is quite complex for machines.
2. Like Elements
   - If elements are similar in features such as shape, colour or size they will people are more ready to consider the data as grouped together
     - This means it can be important to use techniques such as blurring and desaturating visualisations in order to remove or enhance such distinctions as appropriate.
3. Connected Elements
   - Data That is connected in some way (by a line, curve or border) will give the effect of creating clusters or groups in data

Question 9

- 9. Read the slide Effective Encoding of Data in Lecture Note 2 (https://www.lucypark.kr/courses/2015-ba/visualization1.html), if a dataset has quantitative value, what 5 most important attributes or encoding should we consider? Similarly, if a dataset has nominal value, what 5 most important attributes or encoding should we consider?

1. Quantitative Value If data has a quantitative value, the attributes that should be considered are
   (a) Position
   (b) Length
   (c) Angle
   (d) Area
   (e) Volume

2. Nominal Value If data has a nominal value, the five most important attributes to consider are:

   (a) Position of the Data
   (b) Hue
   (c) Texture
   (d) Connection
   (e) Containment

Question 10

- 10. Read the following article: http://www.csc.ncsu.edu/faculty/healey/PP/index.html

(Wk 3) Relational Data Visualisation (Part I)

Refer to:

- Lecture #3
- Tutorial #3

Relational Data Visualisation Part I

Introduction to Graph Visualisation

Tree Visualisation

Connection Approach

Enclosure Approach

Connection+Enclosure Approach

Tree Graphs

Tree graphs show heirarcy so for example:

- Dendrogram from *heirarchical* clustering
- Family tree
- Organisation Mapping plot

Sunburst tree maps are also known as ring charts as shown before at 2a

1. Tree Maps Voronoi Treemaps are very nice

(Wk 4) Relational Data Visualisation (Part I)

Refer to:

- Lecture #4
- Tutorial 4 not yet released.

Exporting HTML Files

When exporting HTML Files with org export the images won't be embedded, you can either:

- use `pandoc --self-contained` and figure out how to make the

mathjax or katex self-contained (or use `mathml`)

- either use pandoc to convert the org file outright
- use pandoc to redo the HTML file
- embed the images as base64
- Make the HTML Stand-alone using pandoc

Making the HTML Standalone

Code something like the following will make the HTML self-contained, but, it will break the references to a degree, making the format less nice.

```
pandoc scratch.html --self-contained -c
  ↪ ~/Dropbox/profiles/Templates/CSS/github-pandoc.css -o scratch2.html
```

Embedding Images

So for svg images I use the following with TEX, I can't quite remember why but I think it was something like `pandoc` doesn't like *TikZ*.

I think I would need to modify this very slightly to accomidate

```
#!/bin/bash

#svgfile=9d9af7b10ca33c3d1bf4e525a1e32af0a2dc5a9a.svg
htmlfile=$1


if [ "$1" == "-h" ]; then
  echo "`basename $0` <SVG> <HTML>"
  exit 0
fi

# for consistency recreate the html file
t2h doseReport.tex github-pandoc.css > t2h.log 2> /dev/null
rm -f t2h.log

ls *svg >> pics.txt
for value in $(cat pics.txt)
do
  svgfile=$value

#Specify the specific Regex to remove
oldtext="<p><img src=\"$svgfile\" \/><\/p>" # This might have been causing issues
#oldtext="$svgfile"

#capture the text of the svg
svtext="$(cat $svgfile)" # This won't work
echo $(sed '' $svgfile) # this won't work either, we
                        #need to escape special characters
newtext=$(sed 's@[/\&]@\\&@g;$!s/$/\\/' $svgfile)
#newtext="=========="
# https://unix.stackexchange.com/a/152192
    #basically the 's@[/\&]@\\&@g;$!s/$/\\/' is necessary to
      #escape all the misbehaved characters


# Identify the line in the html
     # sed -i -e "s/$oldtext/====/g" doseReport.html

# Replace the line
 sed -i -e "s/$oldtext/$newtext/g" $htmlfile

 # Remove the svg
 rm $svgfile
done
rm pics.txt
```

References