Lesson 5:
Visual Multivariate Analysis

AUTHOR
Dr. Kam Tin Seong
Assoc. Professor of Information Systems (Practice)

AFFILIATION
School of Computing and Information Systems,
Singapore Management University

PUBLISHED
07 Feb 2023

Content

What will you learn from this lesson?

- Understand the characteristics of multidimensional data
- Visual analytics techniques and tools for visualising and analysing multidimensional continuous data
- Visual analytics techniques and tools for visualising and analysing multidimensional categorical data
- Sensing both categorical and continuous multidimensional data
- Multidimensional data analysis best practices

In this lesson, I am going to share with you a collection of visual multivariate analysis techniques.

Visual analytics techniques

- Scatterplot Matrix
- Ternary plot
- Glyphs
- Parallel coordinates
- Heatmap

There are many Visual Multivariate Analysis technqiues. In this lesson, I am going share with you five popular one. They are:

Introducing Multidimensional Data

Wine data set

There are 13 variables in this data set. 11 of them are in continuous data type, one in ordinal scale and one in nominal scale.

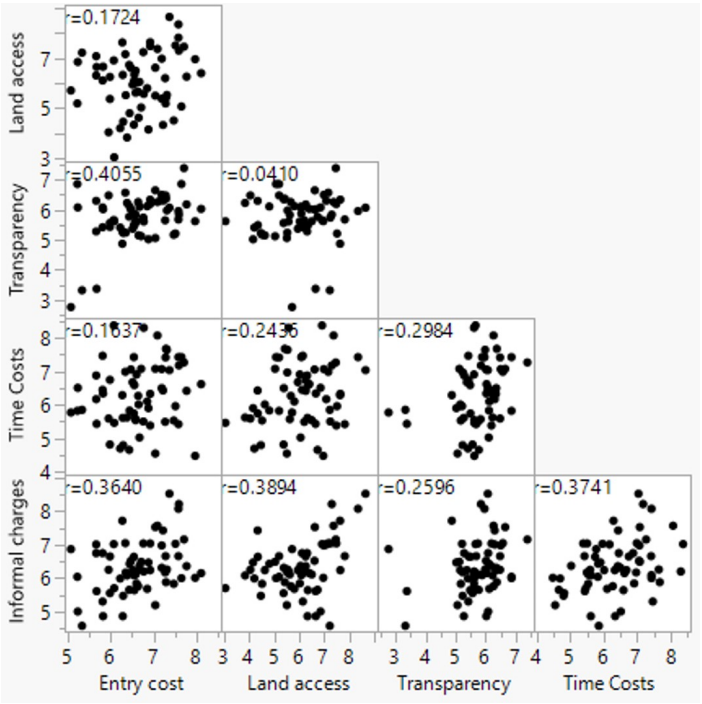
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	red
7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5	red
7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8	5	red
11.2	0.28	0.56	1.9	0.075	17	60	0.9980	3.16	0.58	9.8	6	red
7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	red
7.4	0.66	0.00	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	red

Source: UCI Machine Learning Repository

Table in this slide shows a typical multivariate data sets. This is the popular wine quality data from UCI Machine Learning Repository. It consists of 13 variables in this data set. The first 11 variables are in continuous data type. The quality and type variables are in ordinal scale and nominal scale respectively.

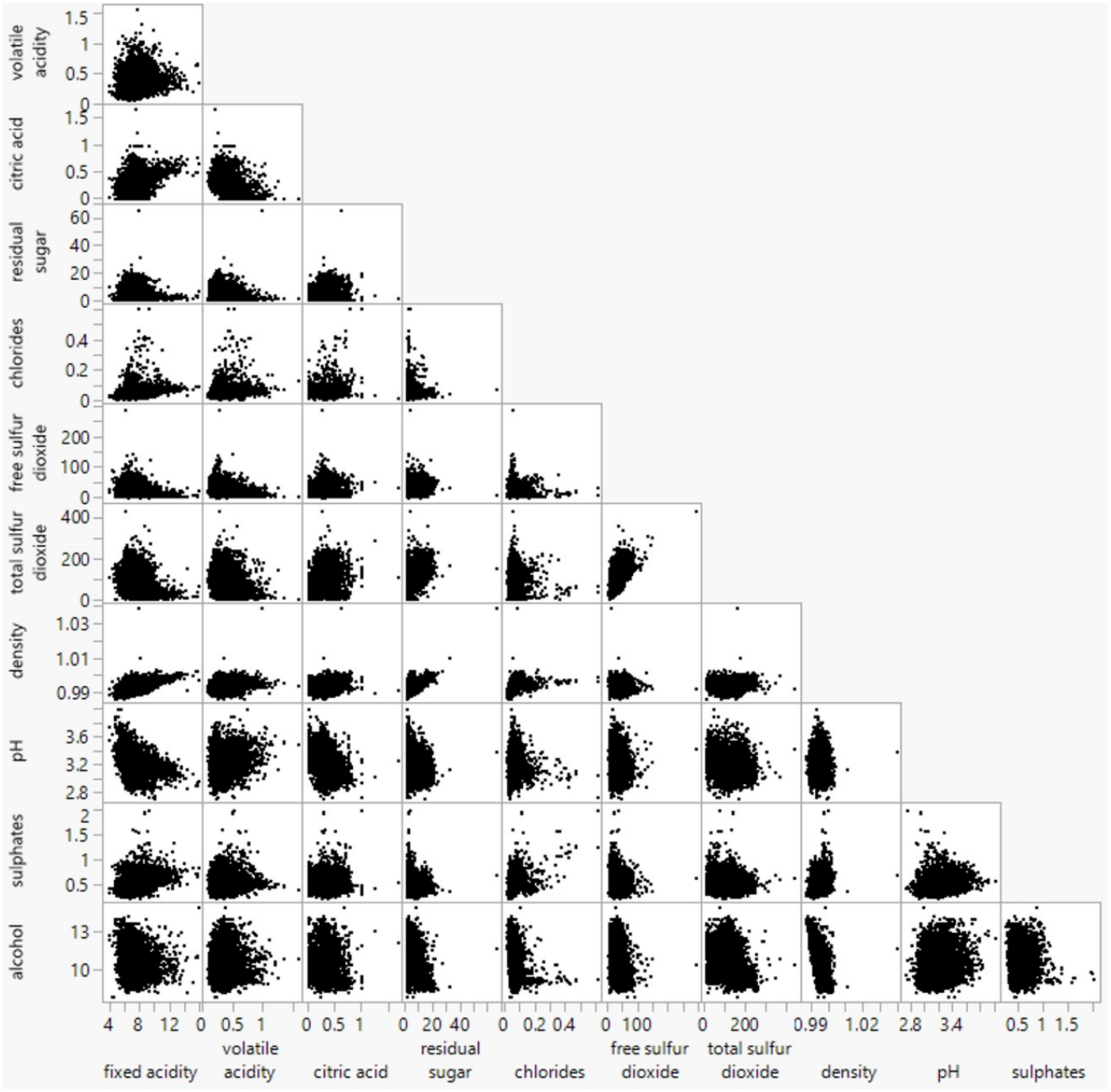
Scatterplot Matrix

- **Scatterplot matrix** (also known as **correlation matrix**) is a graphical method used to reveal the relationship between multiple variables *pairwisely*.



Each black dot in the scatterplot matrix represents an observation in the data table.

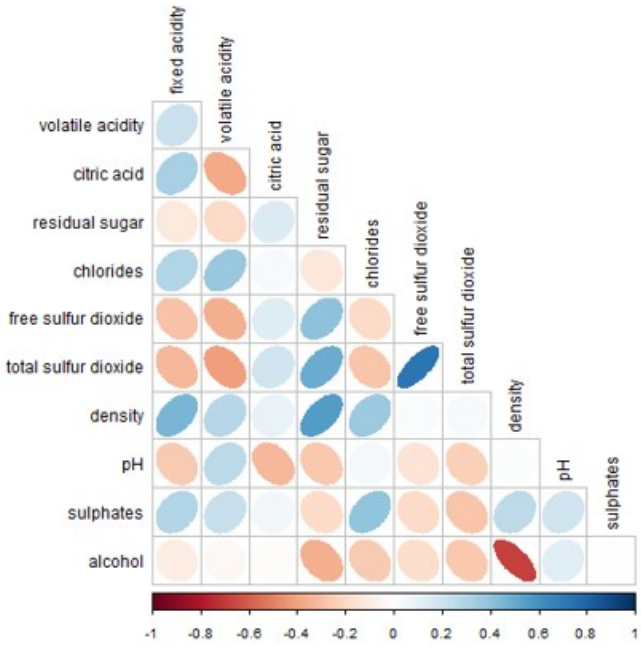
Scatterplot Matrix: Problem with large data



When a relatively large number of observation points is used, the scatterplot matrix failed to reveal the pair relationship effectively.

Correlogram

- `Correlogram` uses **visual geometrics** such as ellipse, circle, square, and bars to replace the scatterplot in correlation matrix.
- It is very useful to reveal pair-relationships between variables in a large correlation matrix.
- In this plot, correlation coefficients is colored according to the value.



For positive correlation, the colour intensity change from light blue to dark blue when the correlation coefficient values increases from 0 to 1. For negative correlation, the colour intensity change from light red to dark red when the correlation coefficient values increase from 0 to -1. Besides colour, the shape and direction of the ellipse also used to map the correlation coefficient. For example, the direction of the ellipses indicate positive or negative correlation and thin ellipses are shown paired variables with strong linear correlation.

Visual abstractions for rendering correlation values.

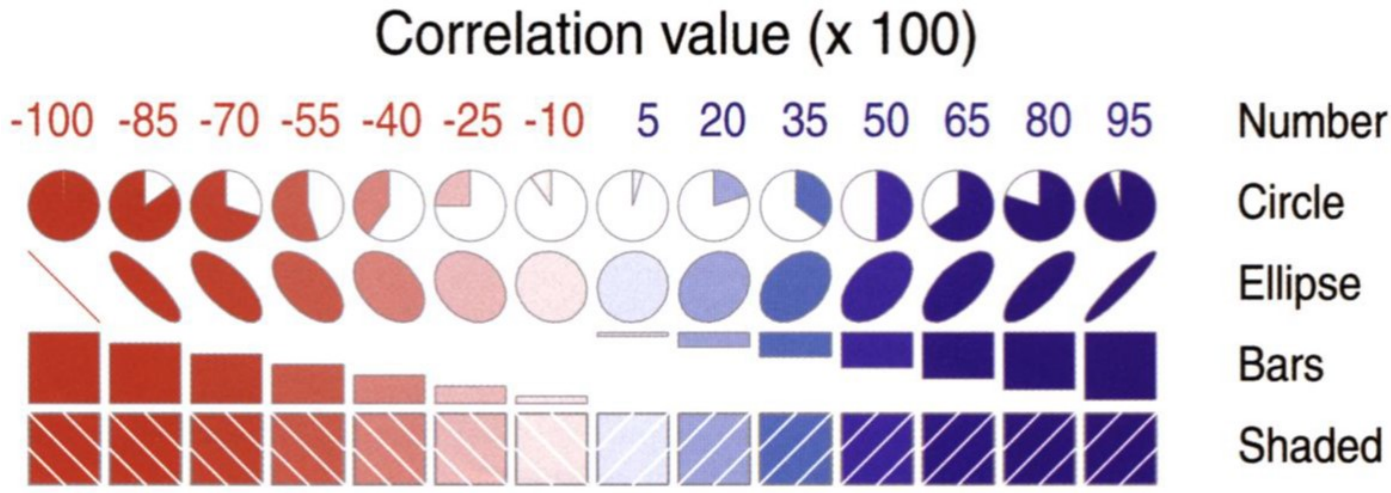
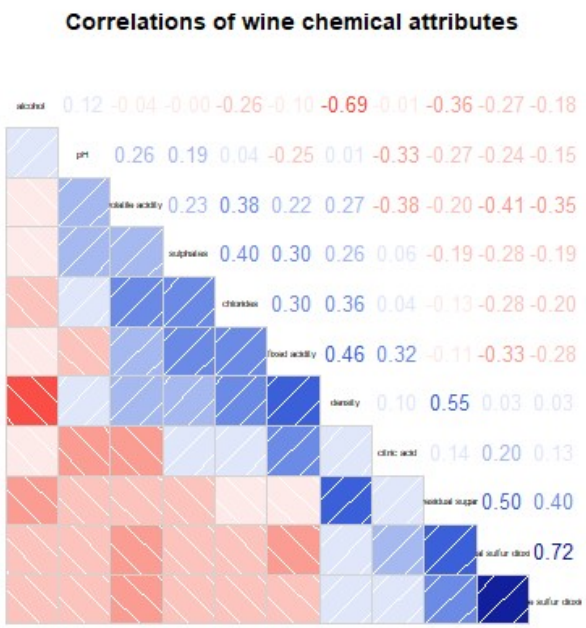


Figure in this slide shows the visual variables commonly used to build correlogram. These include visual objects such as circle, ellipse, bar, and shaded. Other visual variables include colour, pattern, direction and number.

---

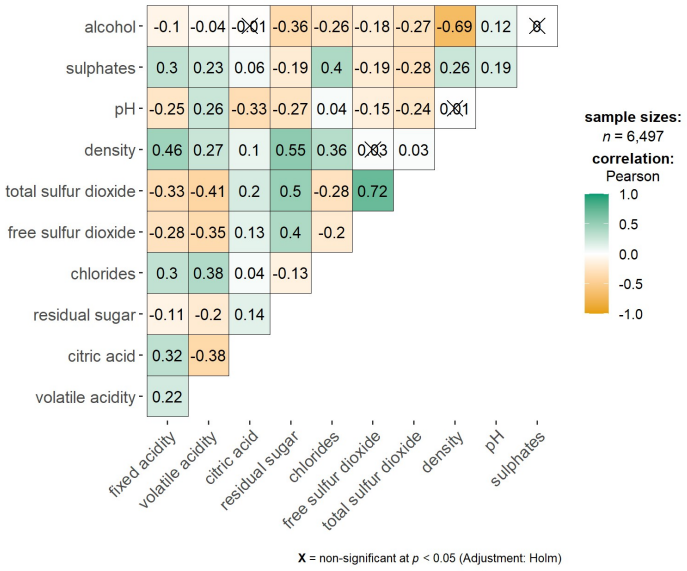
corrgram package

- The corrgram is one of the oldest R package specially designed to correlograms. You can choose what to display in the upper, lower and diagonal part of the figure: scatterplot, pie chart, text, ellipse and more.



ggstatsplot package

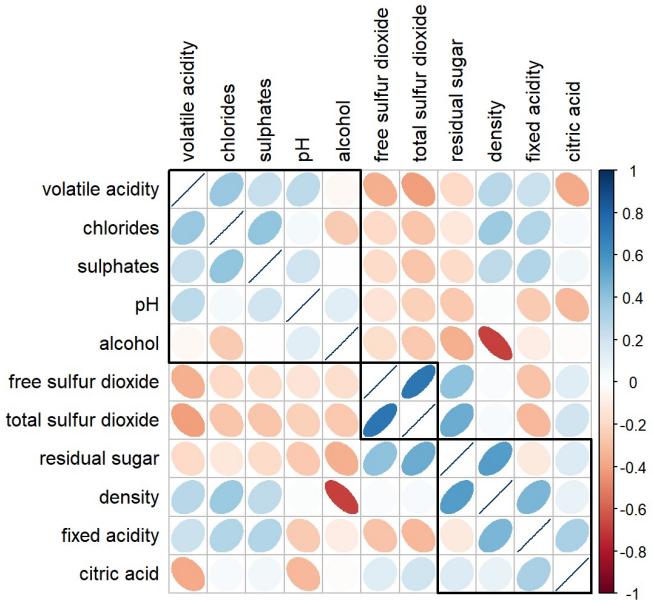
- The *ggcorrmat()* of **ggstatsplot** package can be used to plot a correlogram and their corresponding coefficients.
- To learn more, refer to ggcorrmat and the function's reference guide.



One of the added feature in `ggcorrmat()` of ggstatsplot is the statistics report. The statistical report is conformed to APA standard.
With `ggcorrmat()`, we can have both visual representation and statistical testing report. Furthermore, the statistical test include both frequentist versus bayesian methods.

Corrplot package

- The correlogram on the right is produced by using R **corrplot**.
- The article entitle An Introduction to corrplot Package provides a good overview of the functions of the package.



One of important feature of **corrplot** package that I enjoyed most is it's ability to reorder the correlation matrix by using selected statistical algorithms. Currently, four methods are provided, they are: AOE, FPC, hclust, alphabet. You can find the detail of each methods by visiting the recommended article.

Multivariate data with both continuous and categorical variables

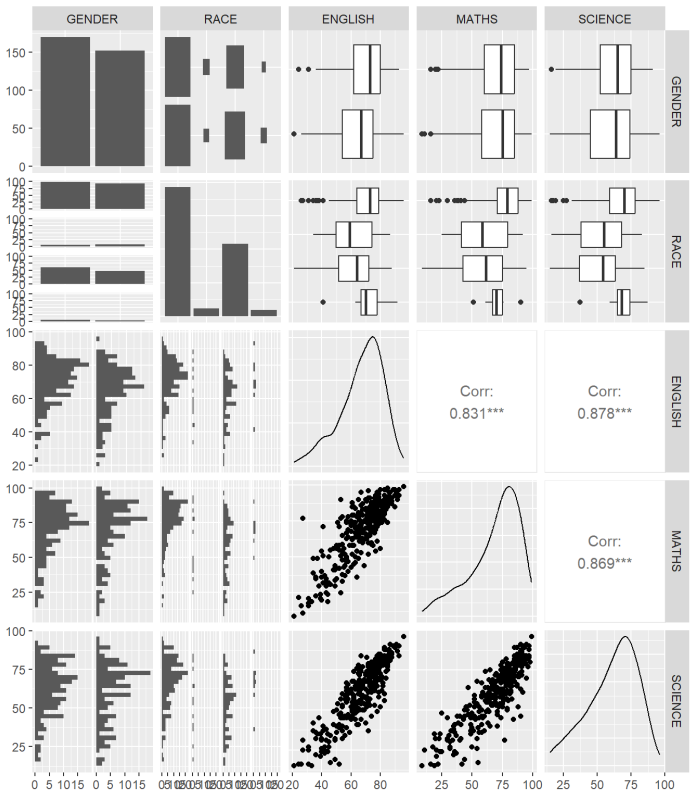
ID	CLASS	TYPE	GENDER	RACE	ENGLISH	MATHS	SCIENCE
Student321	3I		Male	Malay	21	9	15
Student305	3I		Female	Malay	24	22	16
Student289	3H		Male	Chinese	26	16	16
Student227	3F		Male	Chinese	27	77	31
Student318	3I		Male	Malay	27	11	25
Student306	3I		Female	Malay	31	16	16

There are time the data set comprises of both continuous and categorical variables as shown in the slide.

Correlogram: GGally Package

- The GGally package offers great options to build correlograms.
- The *ggpairs()* function build a classic correlogram with scatterplot, correlation coefficient and variable distribution. On top of that, it is possible to inject ggplot2 code, for instance to color categories.
- Visit this link to learn more about Generalised Pairs Plot

Generalised Pairs Plot



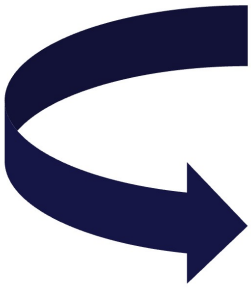
Beyond Visualising Variables Pairwisely

The data

CENSUS OF POPULATION 2010
Table A3 Resident Population by Planning Area, Age Group and Sex

Number

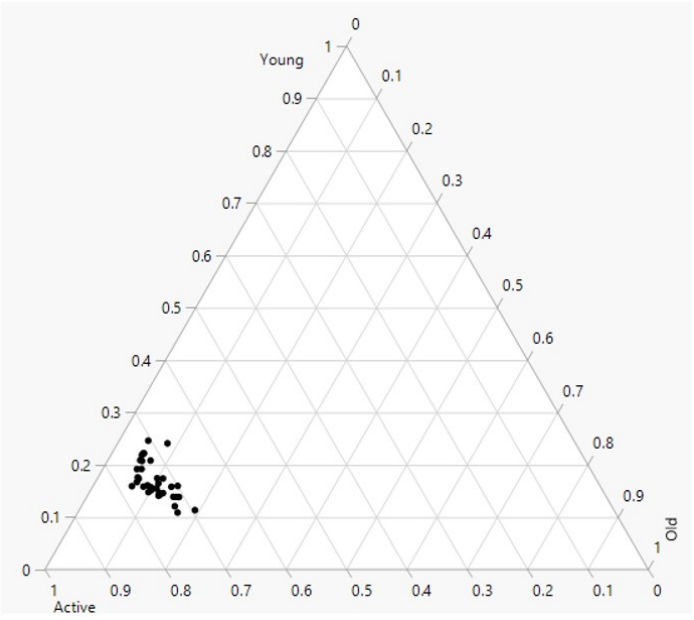
Planning Area	Total	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65 & Over
	Total														
Total	3,771,721	194,432	215,675	244,302	263,750	247,190	272,639	298,687	320,024	309,441	323,459	303,044	248,696	191,995	338,387
Ang Mo Kio	179,297	7,967	8,424	9,335	10,457	10,656	13,400	14,502	14,510	13,525	14,862	14,605	13,785	11,868	21,401
Bedok	294,519	13,230	15,018	17,489	20,083	20,156	21,265	21,707	22,751	22,018	24,615	24,632	21,891	18,018	31,646
Bishan	91,298	3,941	4,759	5,691	7,188	6,338	5,930	6,010	7,354	7,102	8,041	8,245	6,725	5,141	8,833
Bukit Batok	144,198	7,187	8,516	9,738	10,427	10,625	11,332	10,941	11,829	12,259	13,049	12,433	9,911	6,487	9,464
Bukit Merah	157,122	8,049	6,892	6,894	7,479	7,940	10,865	13,871	13,495	11,796	11,650	12,057	11,352	10,782	24,000
Bukit Panjang	128,734	7,106	8,334	9,324	10,224	9,734	9,102	9,488	11,171	10,469	11,528	10,784	7,907	5,316	8,247
Bukit Timah	70,314	3,115	4,456	4,681	5,238	4,613	3,816	4,096	5,689	6,071	6,156	5,726	4,841	4,178	7,638
Changi	2,155	169	192	159	139	94	111	187	252	223	175	122	81	73	178
Choa Chu Kang	173,291	9,232	11,823	15,195	15,362	11,637	11,893	12,541	14,483	16,675	16,919	13,299	9,112	5,855	9,265
Clementi	91,874	4,235	4,567	4,713	4,732	5,133	7,093	7,818	7,831	7,201	7,331	6,882	7,064	6,648	10,626
Downtown Core	3,722	144	122	140	137	161	287	315	326	357	295	295	265	262	616



Planning Zone	Young	Active	Old
Outram	2260	13788	3811
Downtown Core	406	2700	616
Rochor	1899	11341	2424
Queenstown	13777	69670	15055
Bukit Merah	21835	111287	24000
Toa Payoh	17327	88788	18538
Kallang	13861	71438	14260
Marine Parade	7573	33116	6629
Novena	7389	33164	6087
Geylang	17657	88237	14796
Ang Mo Kio	25726	132170	21401
Others	634	3322	528
Clementi	13515	67733	10626
Bukit Timah	12252	50424	7638
Bedok	45737	217136	31646

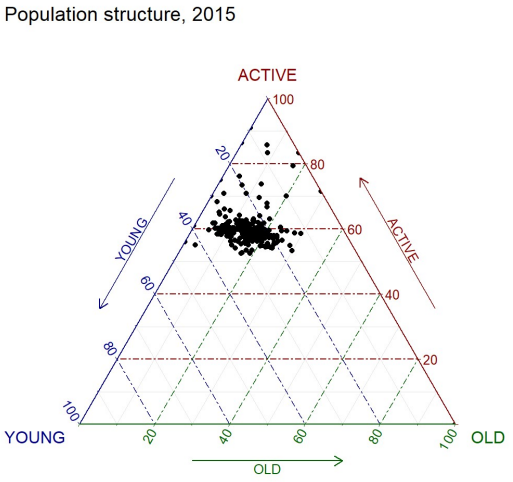
Ternary Plot

- A ternary plot (also known as ternary graph, triangle plot, simplex plot, Gibbs triangle or de Finetti diagram) is a barycentric plot on three variables which sum to a constant, usually in percentage.
- It graphically depicts the ratios of the three variables as positions in an equilateral triangle.



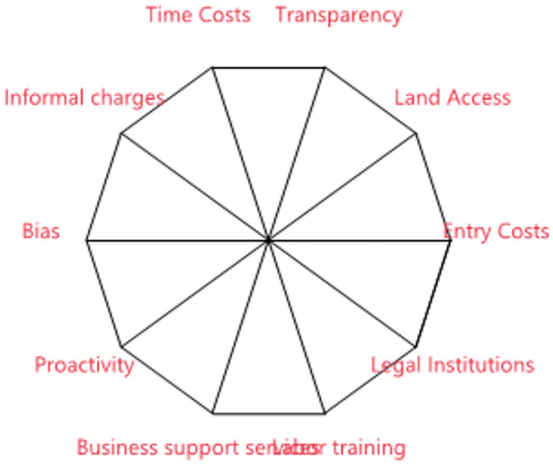
ggtern package

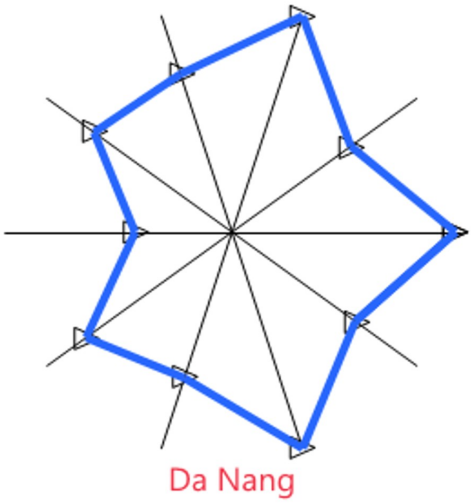
- ggtern is a package that extends the functionality of ggplot2, giving the capability to plot ternary diagrams for (subset of) the ggplot2 proto geometries.
- For a good start, please refer to the article entitle ggtern: Ternary Diagrams Using ggplot2



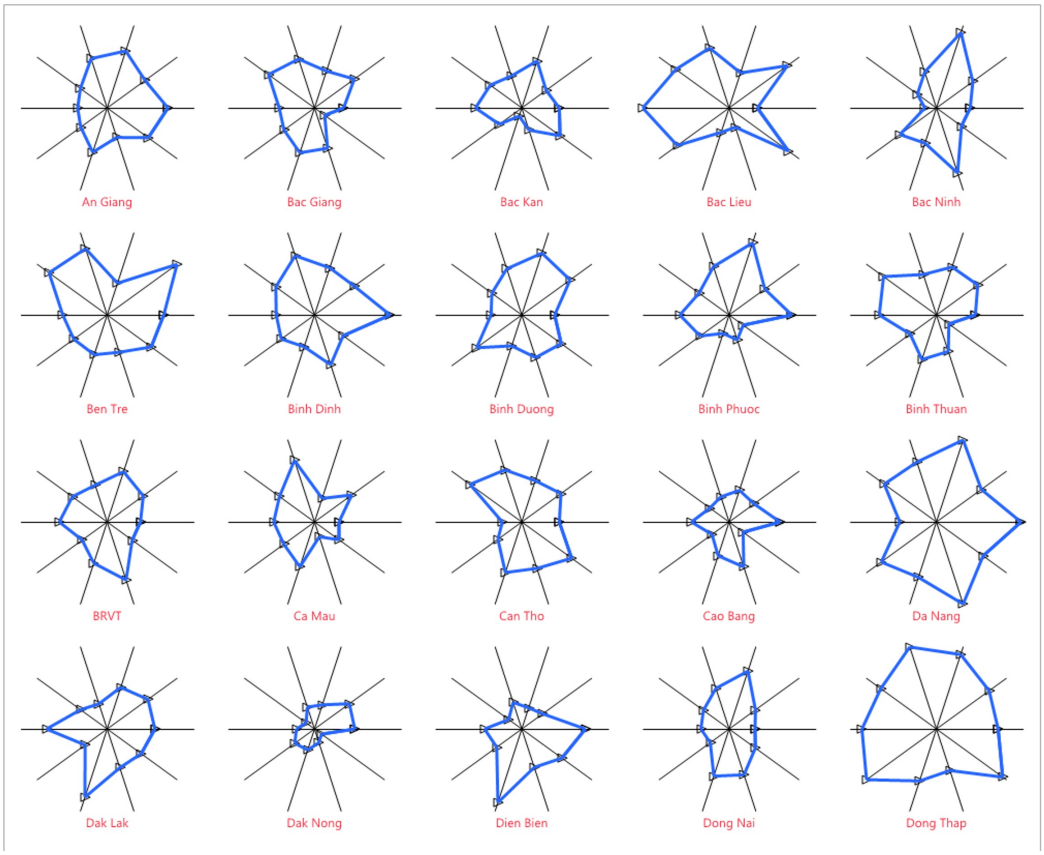
Glyphs

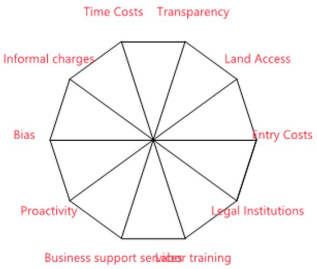
- Star plot (Chambers 1983)(also known as radar chart, star chart and spider chart) is a method of displaying multivariate data.
- The star plot consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables. The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each spoke. This gives the plot a star-like appearance and the origin of the name of this plot.





Multiple Glyphs Chart



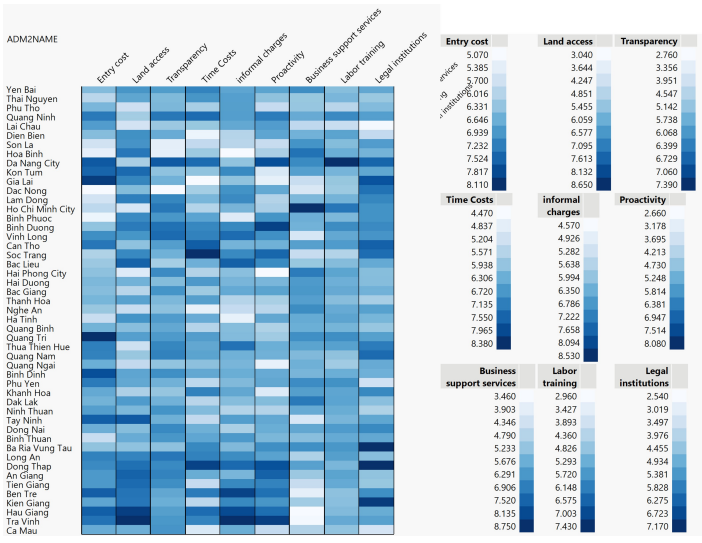


Glyphs Chart in R

- In R, *radarchart()* of **fmsb** library is the best tool to build radar chart.

Visualising and Analysing Multivariate Data: Heatmap method

- A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors.



- When applied to a tabular format, heatmaps are useful for cross-examining multivariate data, through placing variables in the columns and observation (or records) in row and colouring the cells within the table.
- Heatmaps are good for showing variance across multiple variables, revealing any patterns, displaying whether any variables are similar to each other, and for detecting if any correlations exist in-between them.
- It is important to note that when the values are all positive or all negative, a colour scheme with different intensity should be used. On the other hand, if the values are in both positive and negative, then, diverging colour scheme should be used.

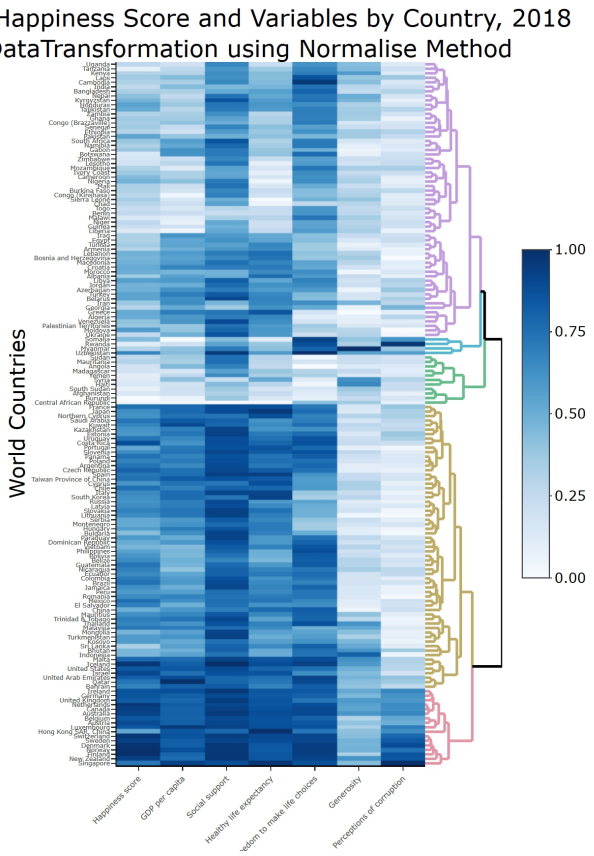
R packages for creating statis Heatmap

There are many R packages and functions can be used to drawing static heatmaps, they are:

- heatmap()of R stats package. It draws a simple heatmap.
- heatmap.2() of **gplots** R package. It draws an enhanced heatmap compared to the R base function.
- pheatmap() of **pheatmap** R package. **pheatmap** package also known as Pretty Heatmap. The package provides functions to draws pretty heatmaps and provides more control to change the appearance of heatmaps.
- **ComplexHeatmap** package of R/Bioconductor package. The package draws, annotates and arranges complex heatmaps (very useful for genomic data analysis). The full reference guide of the package is available here.
- **superheat** package: A Graphical Tool for Exploring Complex Datasets Using Heatmaps. A system for generating extendable and customizable heatmaps for exploring complex datasets, including big data and data with multiple data types. The full reference guide of the package is available here.

R package for creating Interactive Heatmap: heatmaply package

- heatmaply is an R package for building interactive cluster heatmap that can be shared online as a stand-alone HTML file. It is designed and maintained by Tal Galili.
- Before we get started, you should review the Introduction to Heatmaply to have an overall understanding of the features and functions of Heatmaply package. You are also required to have the user manualof the package handy with you for reference purposes.

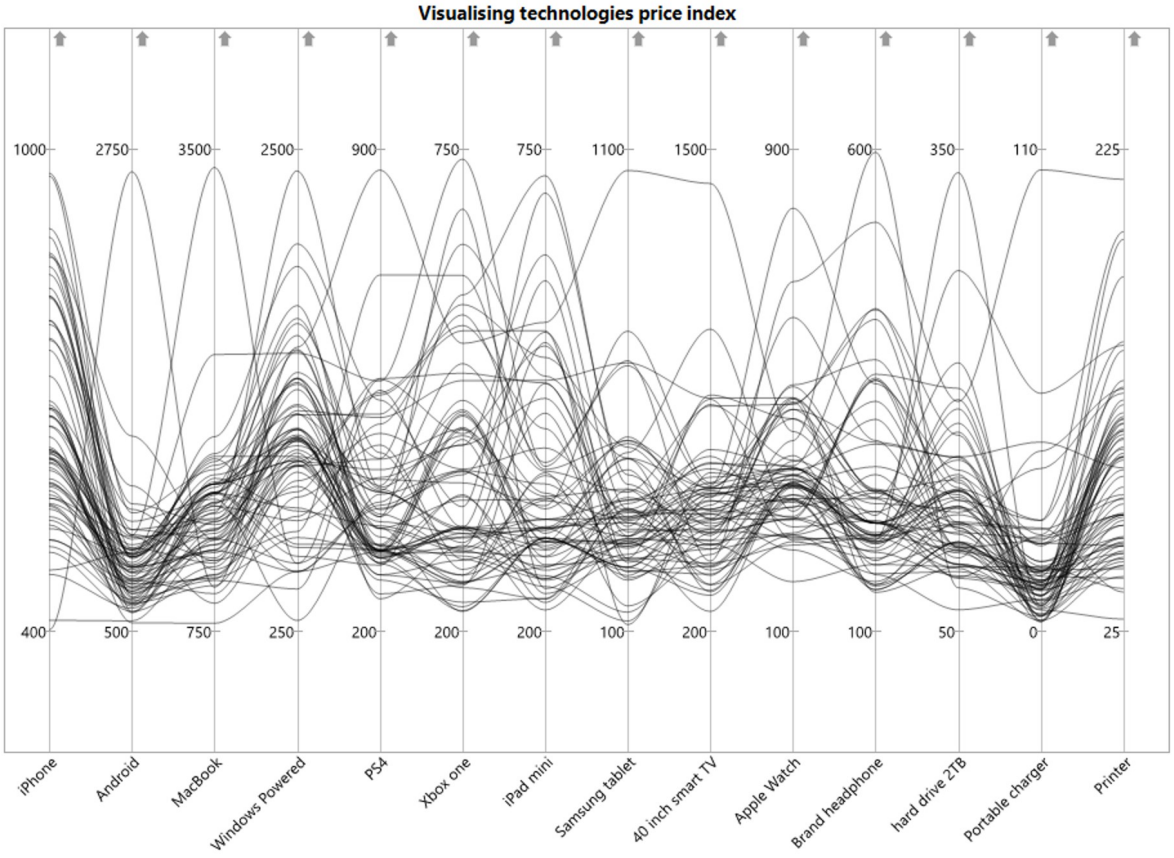


:scale 85%

Visualising and Analysing Multivariate Data

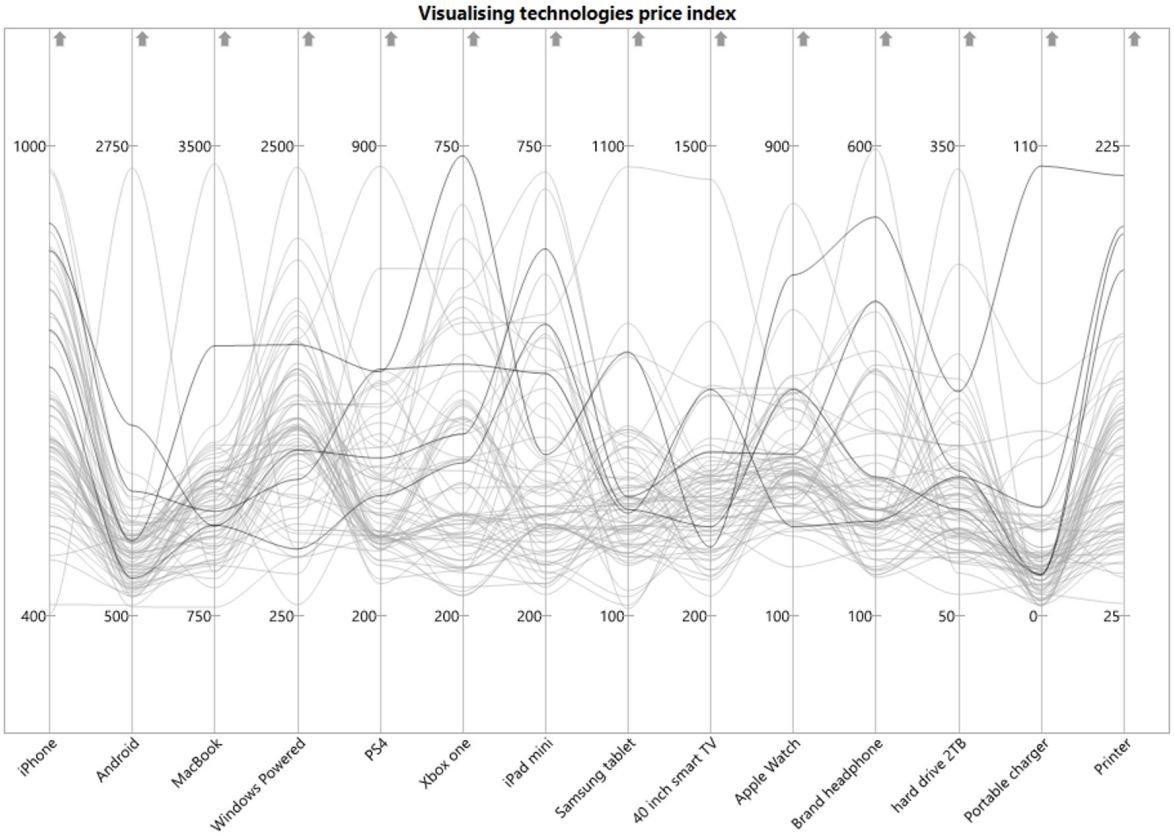
Parallel Coordinates Plot Method

Parallel Coordinates plot or Parallel plot allows to compare the feature of several individual observations on a set of numeric variables.



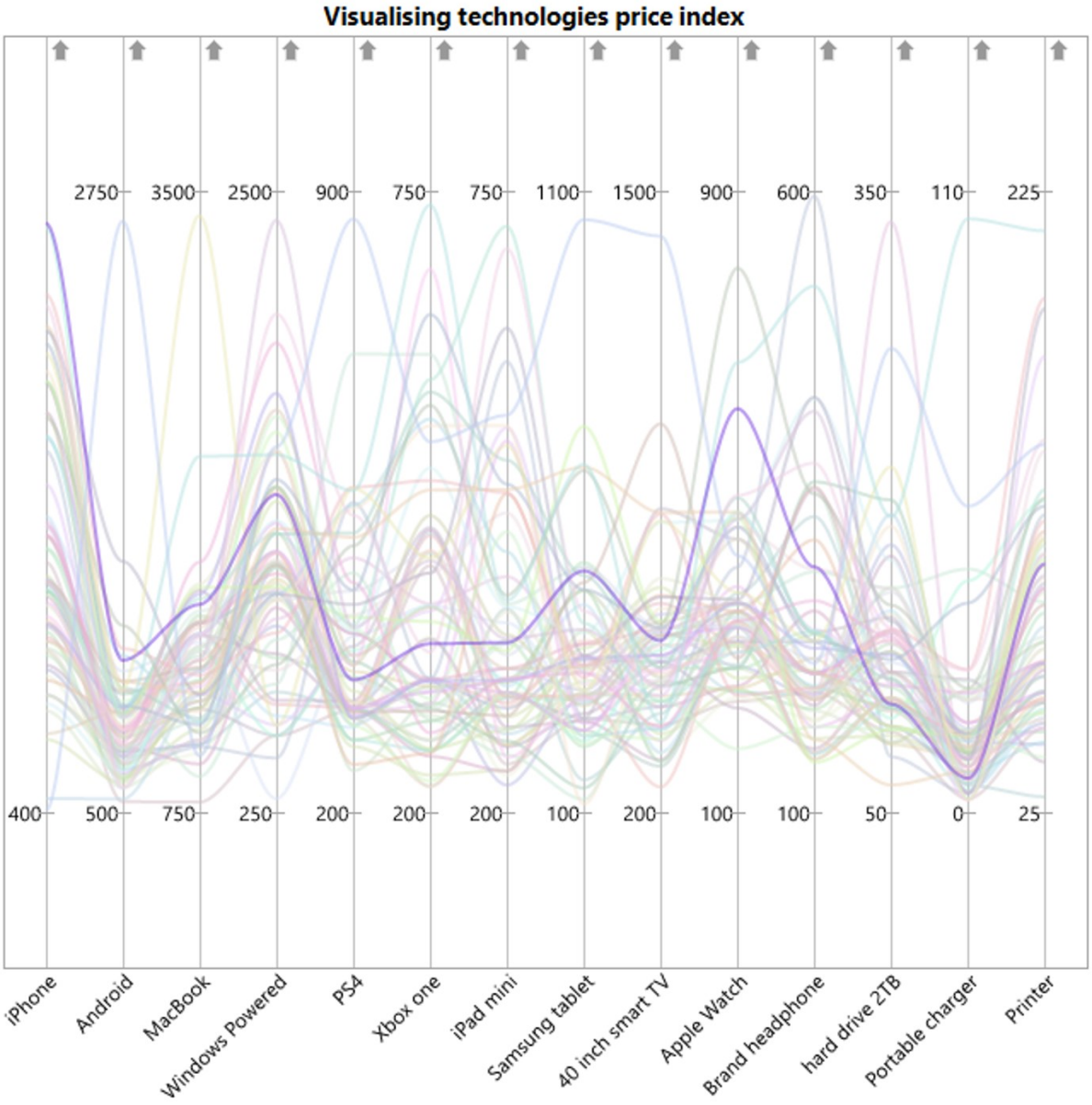
Each vertical bar represents a variable and usually has its own scale. (The units can even be different). Values are then plotted as series of lines connected across each axis. Similar to heatmap, interactivity should be used to complete the data exploration and analysis when using parallel coordinates plot.

Parallel Coordinates: Brushing



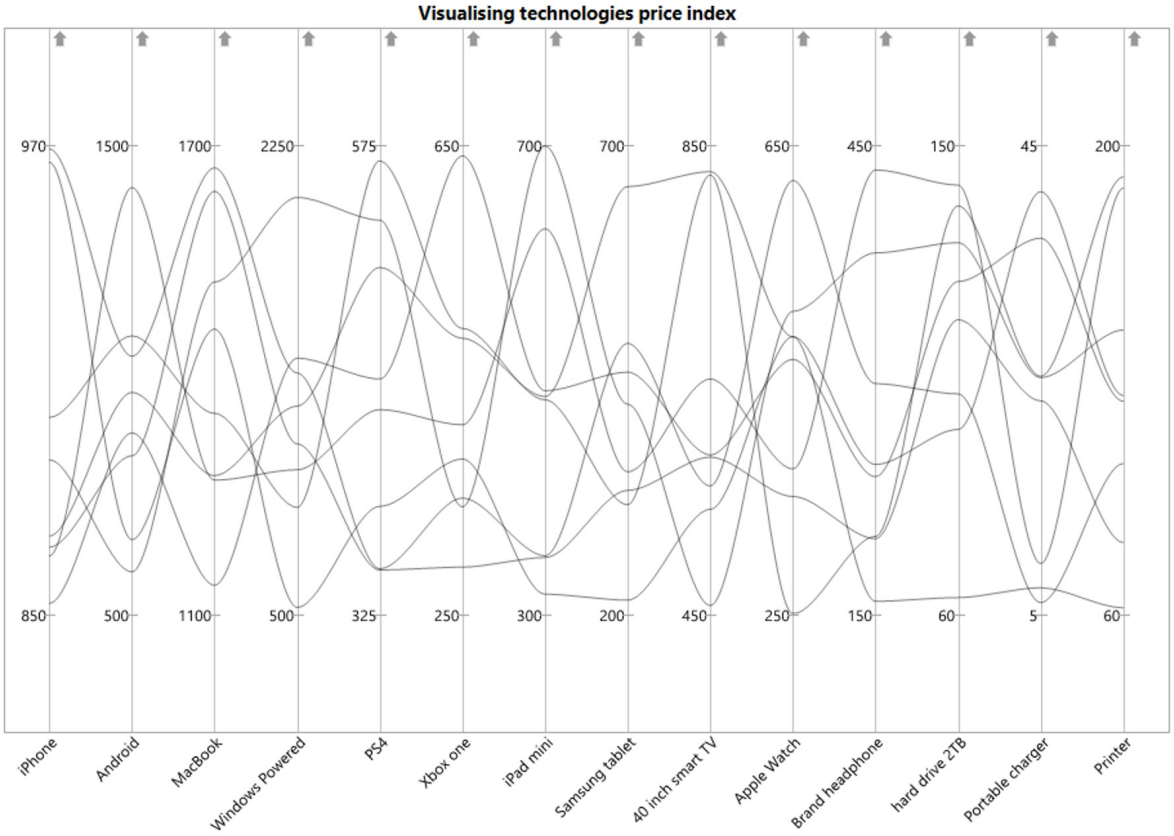
In this slide, brushing is use to select several observations of interest. With brushing function, we can compare them easily.

Parallel Coordinates: Colour and Highlighting



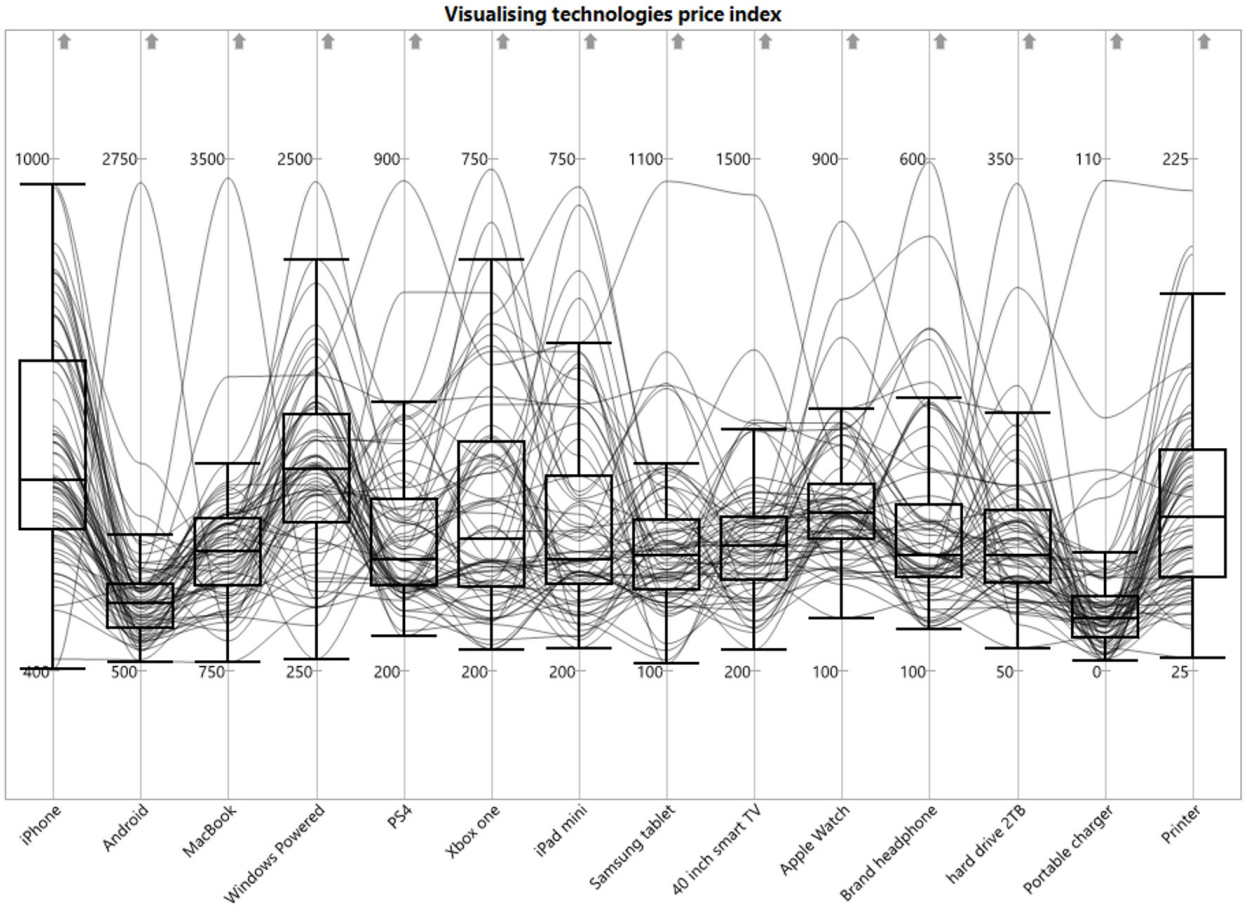
In this slide, colour and highlighting interactive functions is used to select observation of interest.

Parallel Coordinates: Filtering



In this slide, filtering is used to remove unwanted observations so that we can focus on the observation we are interested to investigate.

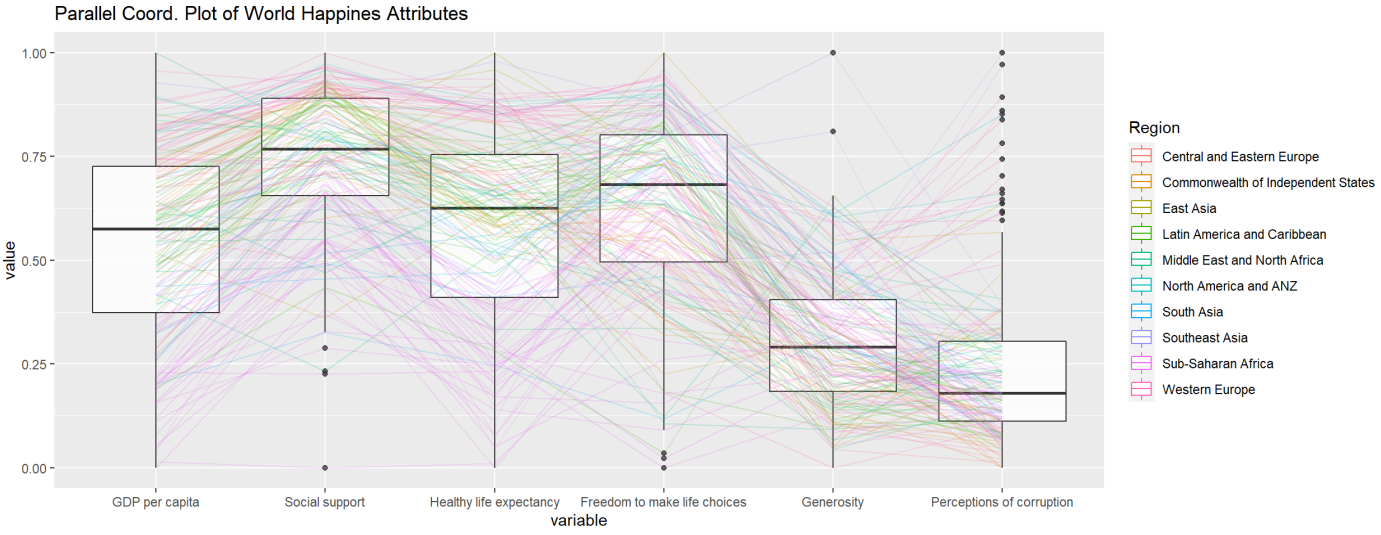
Parallel Coordinates and boxplot



Addition statistical graphics such as boxplot can be used to provide context for more effective visual data exploration and analysis.

Static Parallel Coordinates Plot in R

- ggparcoord() is a function of GGally package for plotting static parallel coordinate plots, utilizing the ggplot2 graphics package.



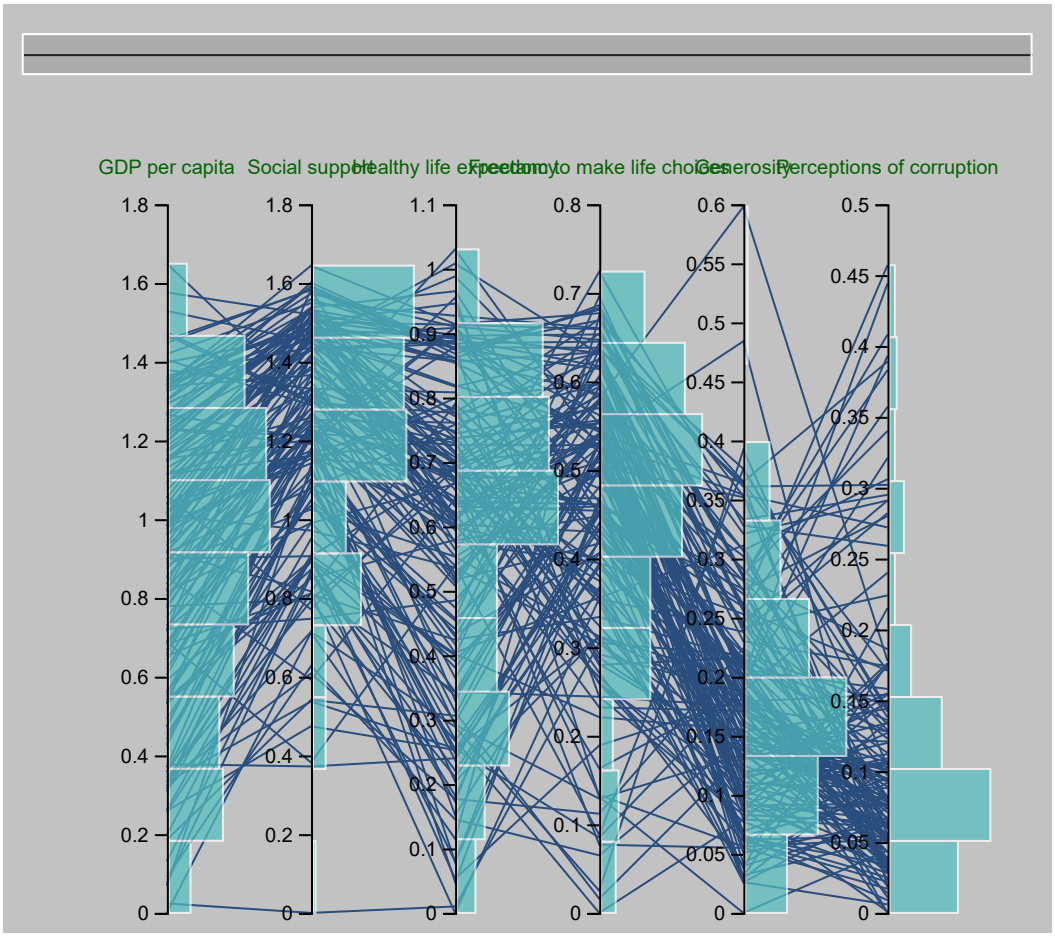
Interactive Parallel Coordinates Plot in R

- **parcoords** package creates interactive parallel coordinates charts with this 'htmlwidget' wrapper for `d3.js`, a JavaScript library for manipulating documents based on data and for creating high interactive data visualisation.

In this slide, brush interactive function is implemented for selecting observation of interest.

Interactive Parallel Coordinates Plot in R

- **parallelPlot** is an R package specially designed to plot a parallel coordinates plot by using 'htmlwidgets' package and d3.js.



In this slide, the colour and highlighting interativity function are used to highlight observation of interest.

Reference

- Radar Chart.
- Ternary plot, this and this.
- Friendly, M (2002) "Corrgrams: Exploratory Displays for Correlation Matrices" The American Statistician, Vol. 56, No. 4, pp. 316-324.