Exploratory Data Analysis (EDA)

PART TWO

Exploratory Data Analysis

Utilize data’s statistical attributes

- Temporal comparison
- Attributes comparison
- Ranking comparison
- Composition analysis
- Distributions analysis
- Variance analysis
- Correlation analysis
- Geographic analysis

Distribution analysis

Compare distribution of attributes

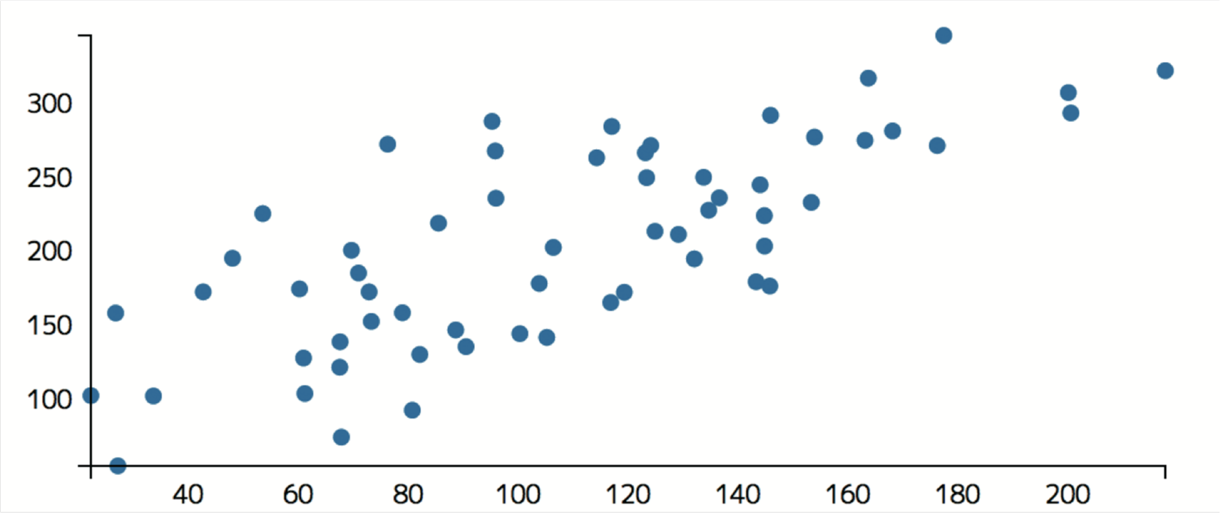
- Visualise /compare
- Clusters
- Spread /distribution

Basic use of various visualisation for Distribution analysis

Basic		Scatter plot
Overall Distribution		Histogram
Distribution of each attributes		Box plot

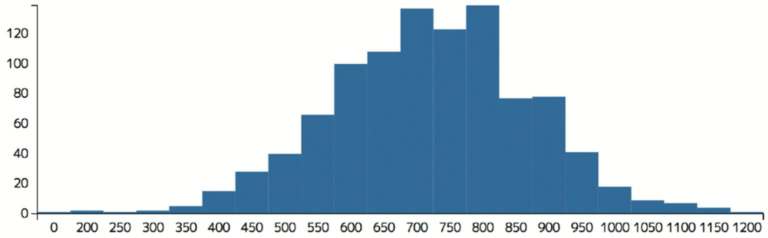
Icons from Data Visualization, Toshikuni and Watanabe (2019)

Scatter plot

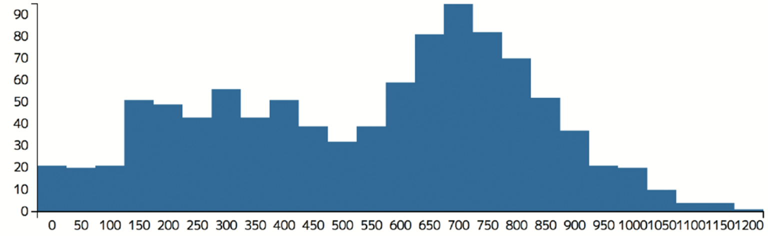


- points

Histogram

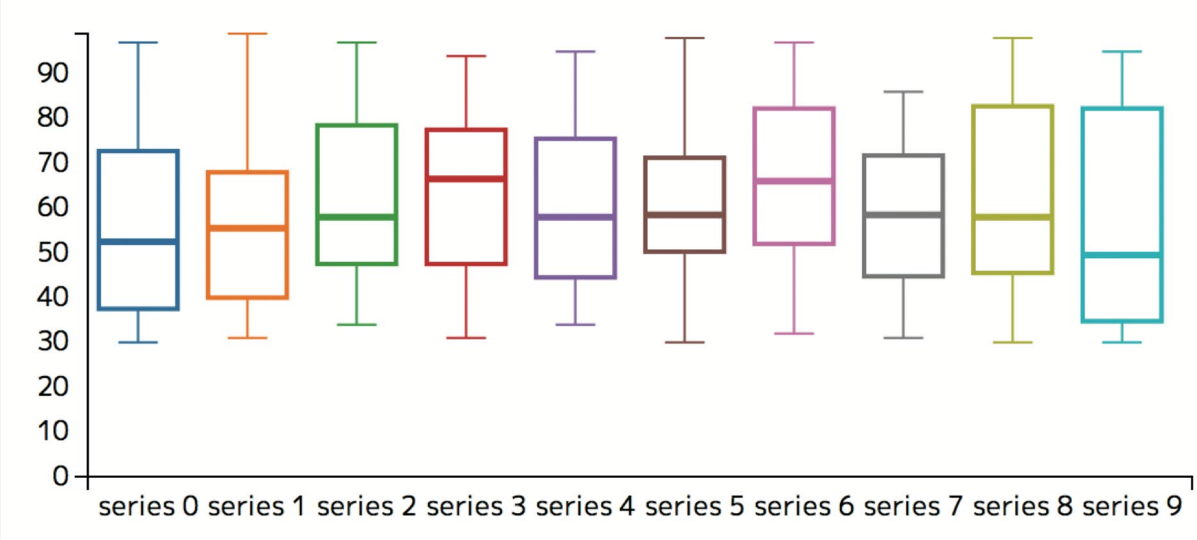


equal space.



- area (width/height)

Box plot



- line, points (3) : median, max, min

Variance Analysis

Compare distribution of attributes with respect to the average

- Visualise /compare
- Clusters
- Spread /distribution

Correlation analysis

Compare distribution of attributes with respect to the average

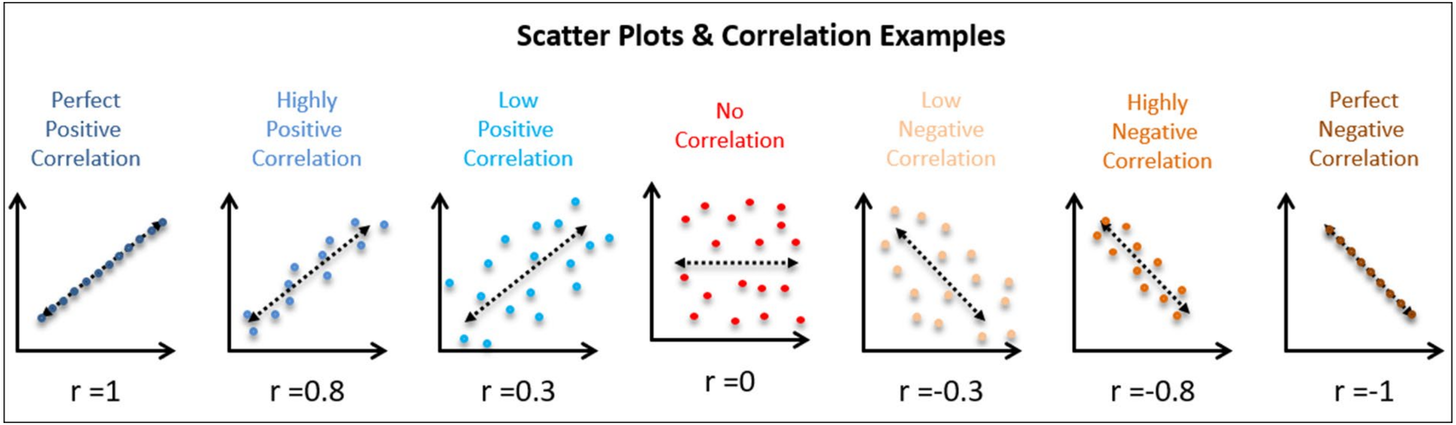
- Visualise /compare
- Positive /negative correlation amount attributes

Basic use of various visualisation for Correlation analysis

Two attributes		Scatter plot
Three attributes		Bubble chart

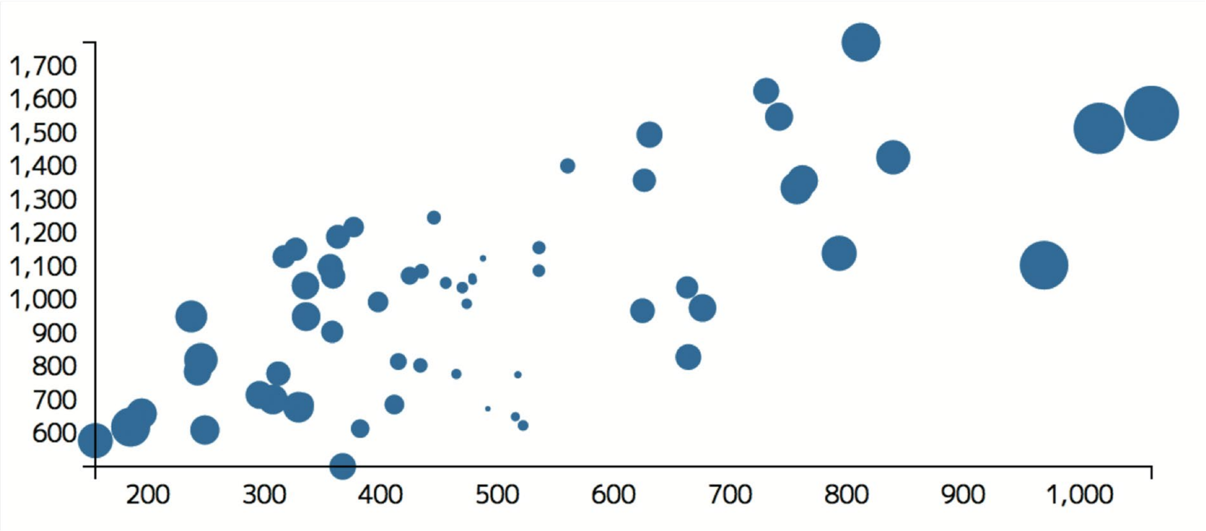
Icons from Data Visualization, Toshikuni and Watanabe (2019)

Correlation Scatter plot



http://www.cqeacademy.com/cqe-body-of-knowledge/continuous-improvement/quality-control-tools/the-scatter-plot-linear-regression/

Bubble chart

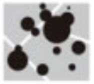


- Circle : value ↔ area

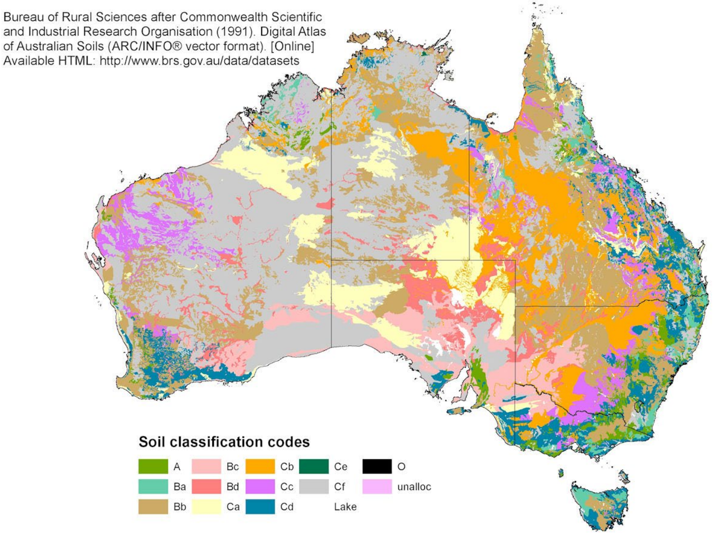
Geographical analysis

- Analyse location/arrangement of attributes based on geo-referenced information

Basic use of various visualisation for Geographical analysis

Compare an index among regions		Thematic map
Compare multiple indices among regions		Symbol map

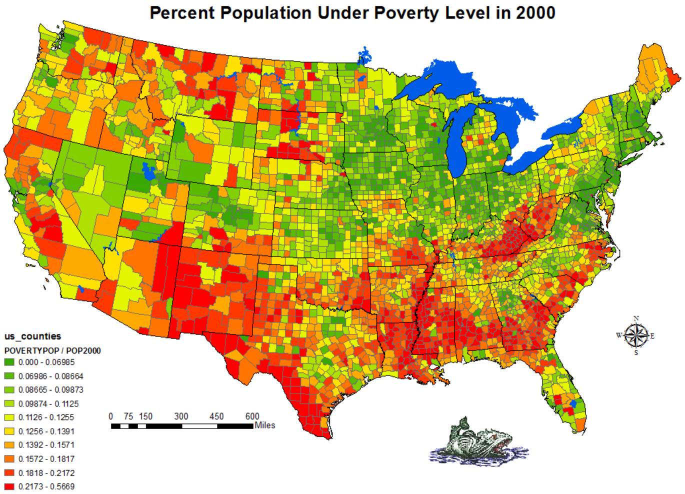
Thematic Map



Bureau of Rural Sciences after Commonwealth Scientific and Industrial Research Organisation (1991). Digital Atlas of Australian Soils (ARC/INFO® vector format). [Online] Available HTML: http://www.brs.gov.au/data/datasets

Soil classification codes

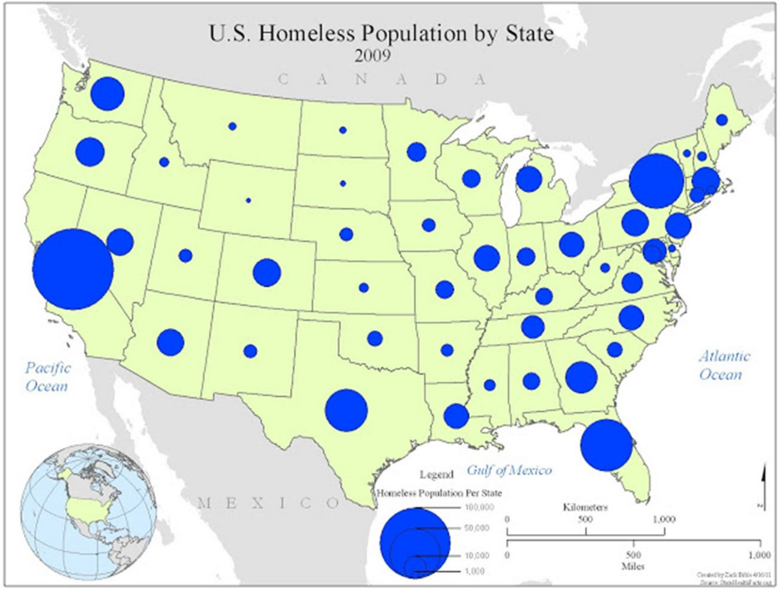
https://www.globalsecurity.org/jhtml/jframe.html#https://www.globalsecurity.org/military/world/australia/images/australia-soils-1.jpg|||



Percent Population Under Poverty Level in 2000

https://mapgeeks.org/different-types-maps/

Symbol Map



http://mph720-jennifer-2012.blogspot.com

Visualising Statistical Feature

Statistics from Data

What sort of statistical feature can you get from the dataset?

- Average,
- Range (minimum and maximum)
- Median
- Variance
- Standard Deviation
- Quartile
- Skewness
- Kurtosis

etc.

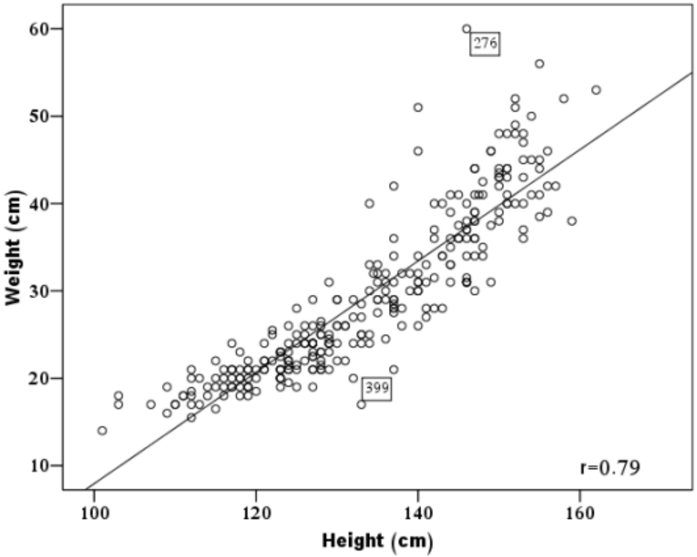
Statistical feature to Visualisation

Statistical values (as indices) can be mapped to graph/chart for visualisation:

- Scatter Plot,
- Histogram
- Probability Plot (Q-Q (quantile-quantile) plot, P-P (Prob-Prov) plot)
- Spaghetti Plot
- Residual Plot
- Box Plot
- Block Plots
- Biplots

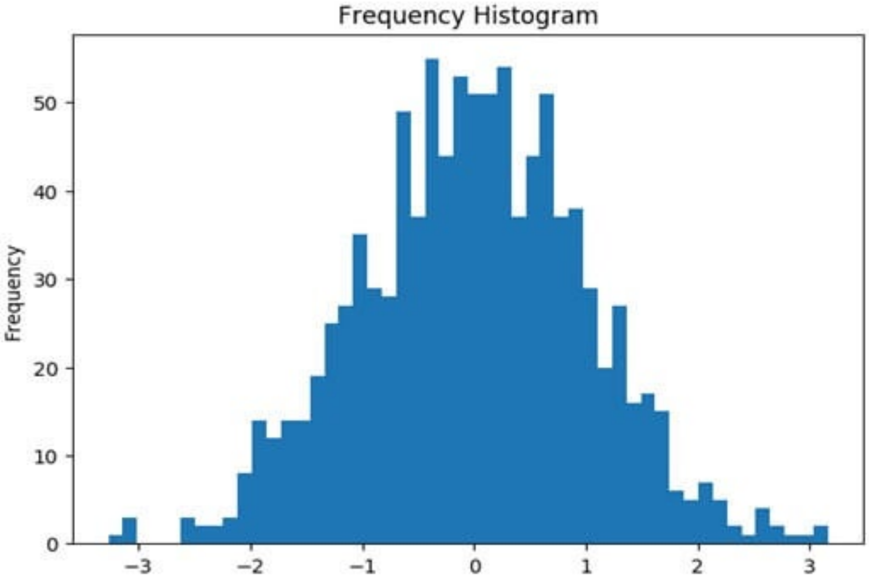
etc.

Scatter Plot



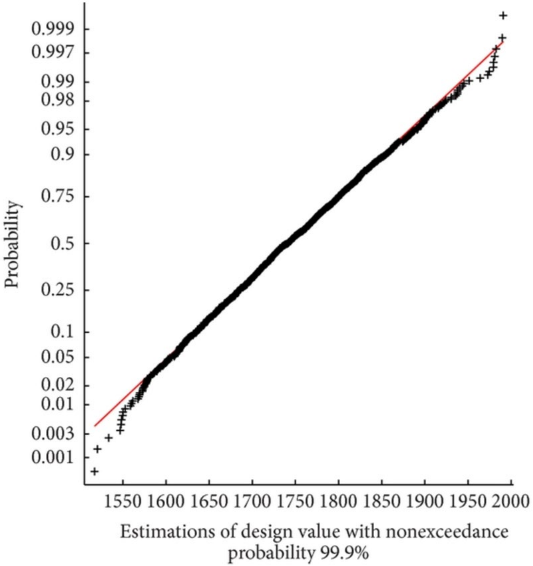
http://www.me-jaa.com/mejaa21Mar2009/scatterplot.htm

Histogram



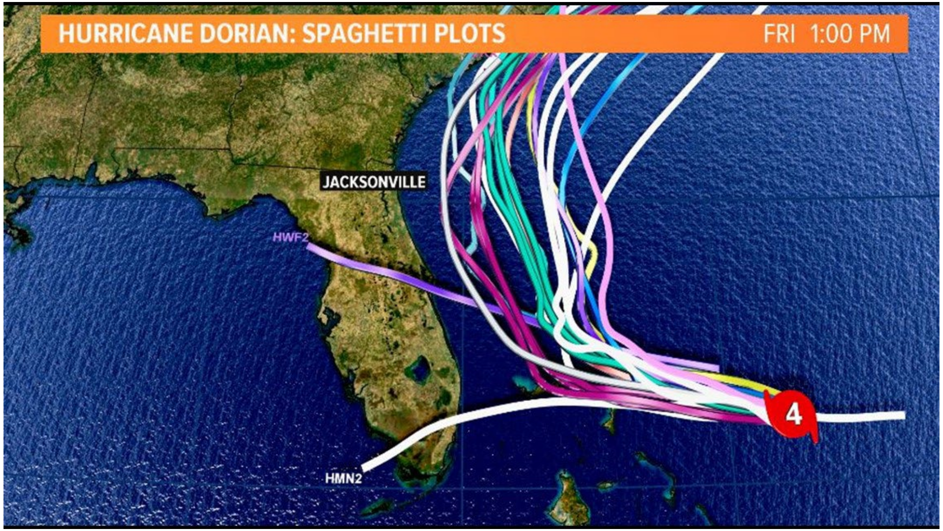
https://pythonclass.in/images/histogram-matplotlib-example.jpg

Probability plot (QQ plot, PP plot)



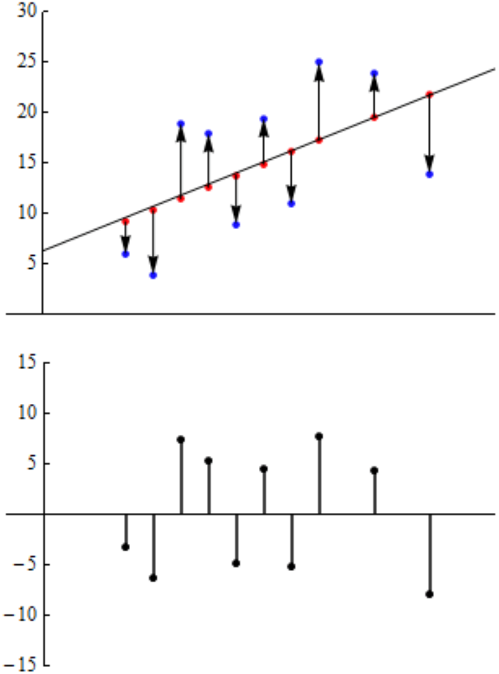
https://www.researchgate.net/profile/Binquan_Li/publication/258397760/figure/download/fig1/AS:319934992273452@1453289854974/Normal-probability-plot-of-estimations-of-the-design-value-xp-with-nonexceedance.png

Spaghetti Plot



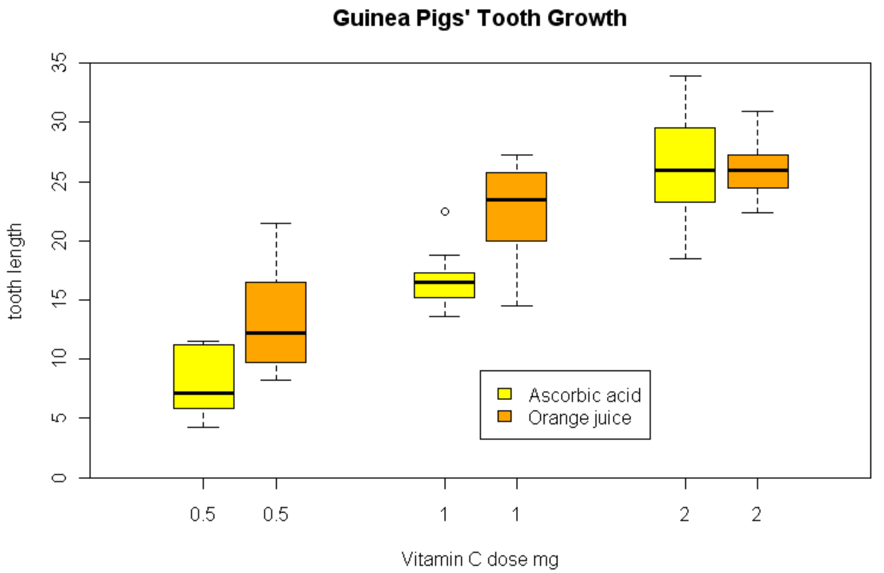
https://media.firstcoastnews.com/assets/WTLV/images/6d45ad7b-9cc4-4e1c-b6b6-efe191dfb3ba/6d45ad7b-9cc4-4e1c-b6b6-efe191dfb3ba_1140x641.jpg

Residual plot



https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/assessing-the-fit-of-a-line-2-of-4/

Box plot

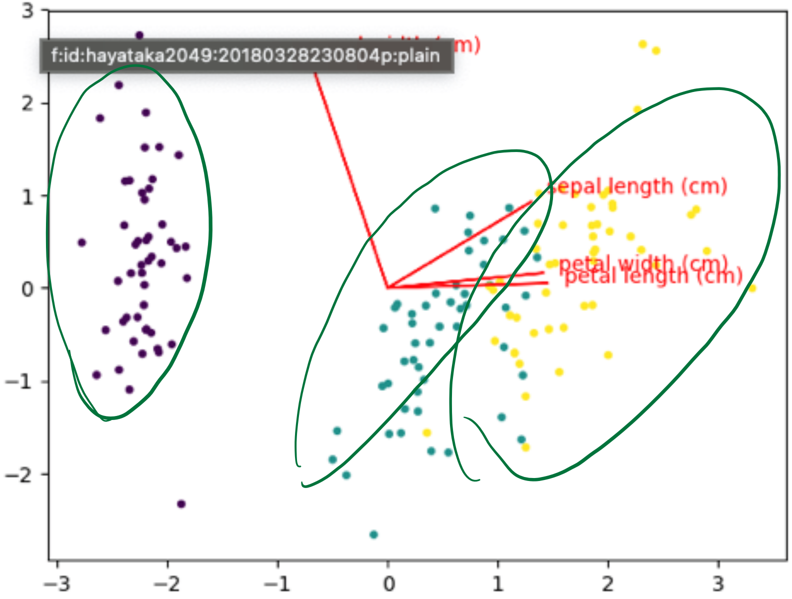


http://4.bp.blogspot.com/_qsPWRbSWL20/TIEKNw3ZziI/AAAAAAAAACw/MIgJoc4Mnvo/s1600/boxplot.PNG

Biplots (PCA)

how clusters related in original space

new dimension A



new dimension B

https://www.haya-programming.com/entry/2018/03/28/231305

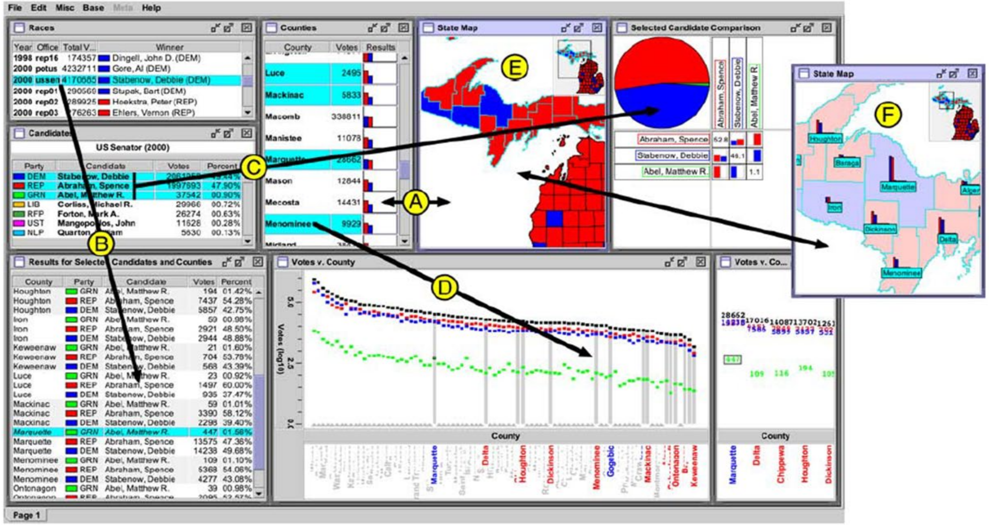
Multi-dimensional Statistical Features Handling

How do we handle multi-dimensional statistical features?

- Coordinated Multiview Visualization
- Spatialisation

Coordinated Multiview Visualization

- A typical visualization used to display statistical features are 2D, or 3D.



Races

Year	Office	Total V...	Winner
1998	rep16	174357	🟦 Dingell, John D. (DEM)
2000	potus	4232711	🟦 Gore, Al (DEM)
2000	ussen	4170685	🟦 Stabenow, Debbie (DEM)
2000	rep01	290569	🟦 Stupak, Bart (DEM)
2000	rep02	289925	🟥 Hoekstra, Peter (REP)
2000	rep03	276263	🟥 Ehlers, Vernon (REP)

Candidates

US Senator (2000)

Party	Candidate	Votes	Percent
🟦 DEM	Stabenow, Debbie	[illegible]	[illegible]
🟥 REP	Abraham, Spence	1997693	47.90%
🟩 GRN	Abel, Matthew R.	37542	00.90%
🟨 LIB	Corliss, Michael R.	29966	00.72%
🟩 RFP	Forton, Mark A.	26274	00.63%
🟪 UST	Mangopoulos, John	11628	00.28%
🟦 NLP	Quarton, William	5630	00.13%

Counties

County	Votes	Results
Luce	2495	
Mackinac	5833	
Macomb	338811	
Manistee	11078	
Marquette	28662	
Mason	12644	
Mecosta	14431	
Menominee	9929	
Midland	[illegible]	

Results for Selected Candidates and Counties

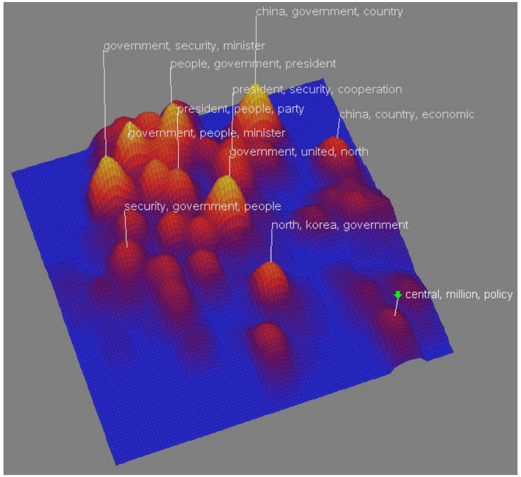
County	Party	Candidate	Votes	Percent
Houghton	🟩 GRN	Abel, Matthew R.	194	01.42%
Houghton	🟥 REP	Abraham, Spence	7437	54.28%
Houghton	🟦 DEM	Stabenow, Debbie	5857	42.75%
Iron	🟩 GRN	Abel, Matthew R.	59	00.98%
Iron	🟥 REP	Abraham, Spence	2921	48.50%
Iron	🟦 DEM	Stabenow, Debbie	2944	48.88%
Keweenaw	🟩 GRN	Abel, Matthew R.	21	01.60%
Keweenaw	🟥 REP	Abraham, Spence	704	53.78%
Keweenaw	🟦 DEM	Stabenow, Debbie	568	43.39%
Luce	🟩 GRN	Abel, Matthew R.	23	00.92%
Luce	🟥 REP	Abraham, Spence	1497	60.00%
Luce	🟦 DEM	Stabenow, Debbie	935	37.47%
Mackinac	🟩 GRN	Abel, Matthew R.	59	01.01%
Mackinac	🟥 REP	Abraham, Spence	3390	58.12%
Mackinac	🟦 DEM	Stabenow, Debbie	2298	39.40%
Marquette	🟩 GRN	Abel, Matthew R.	447	01.56%
Marquette	🟥 REP	Abraham, Spence	13575	47.36%
Marquette	🟦 DEM	Stabenow, Debbie	14238	49.68%
Menominee	🟩 GRN	Abel, Matthew R.	109	01.10%
Menominee	🟥 REP	Abraham, Spence	5368	54.06%
Menominee	🟦 DEM	Stabenow, Debbie	4277	43.08%
Ontonagon	🟩 GRN	Abel, Matthew R.	39	00.98%
Ontonagon	🟥 REP	Abraham, Spence	2095	52.57%

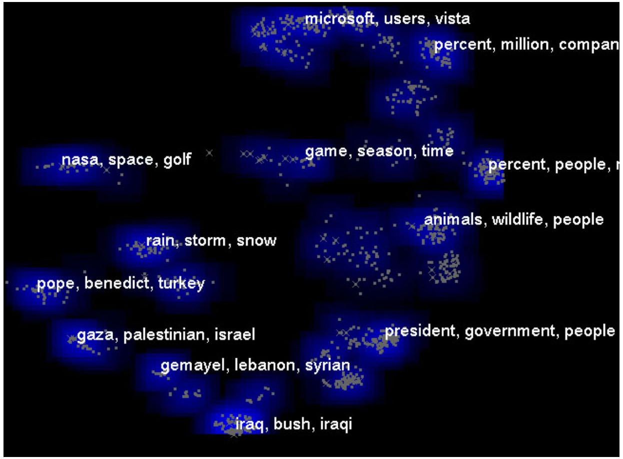
Figure 11: Visualization of election results in Michigan from 1998 to 2002. (A) Shared selection of counties between a table view and a map. (B) Selecting a race causes the election results for that race to be loaded (from a file) and shown throughout the visualization. (C) A pie chart uses a filter to compare results for selected candidates only. (D) A scatterplot highlights selected counties with gray bars. (E) A four-layer scatterplot colors counties by winning candidate party. (F) Semantic zoom labels counties with nested bar plots at sufficient zoom.

Weaver, C. (2004, 10-12 Oct. 2004). Building Highly-Coordinated Visualizations in Improvise. IEEE Symposium on Information Visualization,

Spatialisation

- Mapping multidimensional space to a lower dimensional space


china, government, country
government, security, minister
people, government, president
president, security, cooperation
president, people, party
china, country, economic
government, people, minister
government, united, north
security, government, people
north, korea, government
central, million, policy



Summary

- Exploratory Visual Data Analysis : try to understand the data through visualization of various statistical data.

- Depending on what sort of analysis you would like to carry out, you should choose appropriate visualization.

- Statistics that can be used for visualization
  - Various statistical features can be mapped to visual attributes to assist EDA processes.

- Other techniques used in EDA
  - Various multi-dimensional scaling methods can be used to place multi-dimensional data point on the 2D/3D space to study the complex data.

THE UNIVERSITY OF
SYDNEY