Data Exploration and Visualisation
Part 2: Data Visualisation

Dr Tom Diethe

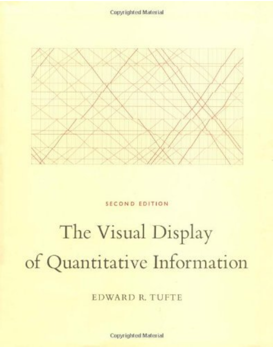
tom.diethe@bristol.ac.uk

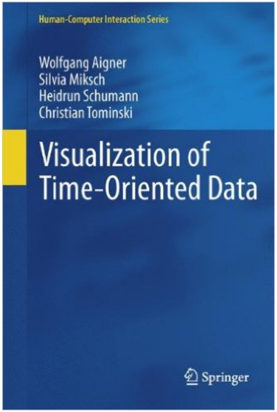
University of Bristol

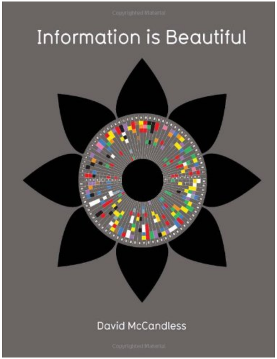
Recap and Outlook

Outline

Resources


Copyrighted Material
SECOND EDITION
The Visual Display
of Quantitative Information
EDWARD R. TUFTE
Copyrighted Material


Human-Computer Interaction Series
Wolfgang Aigner
Silvia Miksch
Heidrun Schumann
Christian Tominski
Visualization of
Time-Oriented Data
Springer


Information is Beautiful
David McCandless

Outline

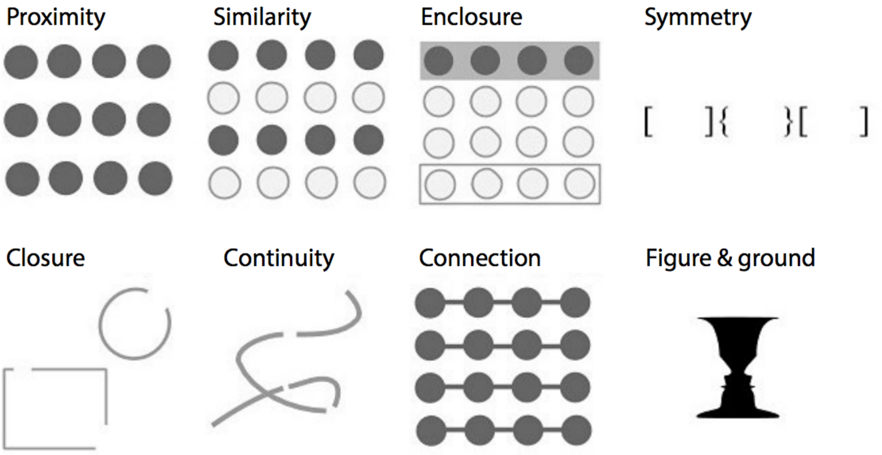
Why Visualise?

- Making sense of data
- Discovery
- Communication
- Monitoring / Situational awareness

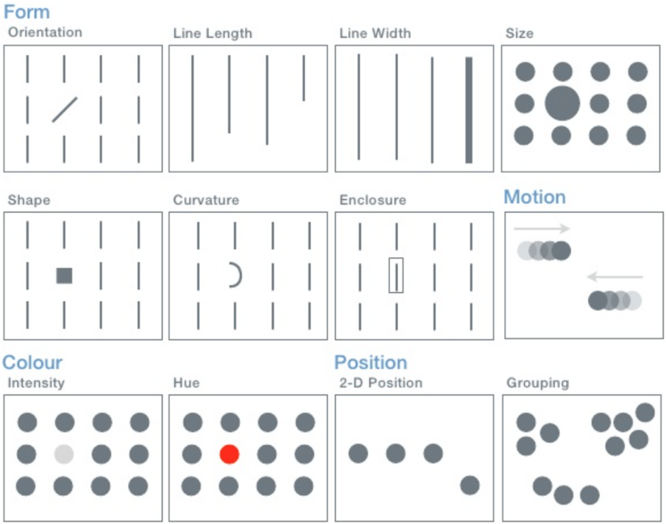
- We can detect information faster than we can move
- Humans are not very good at detecting patterns from numbers

Outline

Gestalt Principles (Bruce et al., 2003)



Preattentive Features



- Criteria for evaluation:
  - Which design minimises eye travel?
  - Which design looks best as black and white? (or colourblind)
  - Maximise information to ink ratio
- What to choose when
  - Line graph: to track changes over periods of time
  - Pie Chart: (nearly) never!
  - Divided Rectangle (Waffle): when you are trying to compare parts of a whole
  - Bar Graph: to compare things between different groups
  - Histogram: to track changes over time, or probability distributions. Note with histograms, the width is significant, as well as the height, unlike a bar graph

- Criteria for evaluation:
  - Which design minimises eye travel?
  - Which design looks best as black and white? (or colourblind)
  - Maximise information to ink ratio
- What to choose when
  - Line graph: to track changes over periods of time
  - Pie Chart: (nearly) never!
  - Divided Rectangle (Waffle): when you are trying to compare parts of a whole
  - Bar Graph: to compare things between different groups
  - Histogram: to track changes over time, or probability distributions. Note with histograms, the width is significant, as well as the height, unlike a bar graph

Age	1980	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003
< 15	607	624	578	523	502	492	479	498	497	519	537
$15 - 19$	451	462	449	418	379	364	347	346	337	341	337
$20 - 24$	310	328	327	328	333	326	314	301	296	298	293
$25 - 29$	213	219	216	213	224	230	224	226	221	219	211
$30 - 34$	213	203	197	189	192	189	179	176	171	171	167
$35 - 39$	317	280	265	244	241	234	219	208	200	195	186
≥ 40	461	409	374	350	339	329	309	291	283	276	268

- Which group has the highest/lowest rates? When?
- Which group has an increasing/decreasing temporal trend?
- Which group has a faster/slower rate of change?

Age	1980	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003
< 15	607	624	578	523	502	492	479	498	497	519	537
$15 - 19$	451	462	449	418	379	364	347	346	337	341	337
$20 - 24$	310	328	327	328	333	326	314	301	296	298	293
$25 - 29$	213	219	216	213	224	230	224	226	221	219	211
$30 - 34$	213	203	197	189	192	189	179	176	171	171	167
$35 - 39$	317	280	265	244	241	234	219	208	200	195	186
≥ 40	461	409	374	350	339	329	309	291	283	276	268

- Which group has the highest/lowest rates? When?
- Which group has an increasing/decreasing temporal trend?
- Which group has a faster/slower rate of change?

Age	1980	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003
< 15	607	624	578	523	502	492	479	498	497	519	537
$15 - 19$	451	462	449	418	379	364	347	346	337	341	337
$20 - 24$	310	328	327	328	333	326	314	301	296	298	293
$25 - 29$	213	219	216	213	224	230	224	226	221	219	211
$30 - 34$	213	203	197	189	192	189	179	176	171	171	167
$35 - 39$	317	280	265	244	241	234	219	208	200	195	186
≥ 40	461	409	374	350	339	329	309	291	283	276	268

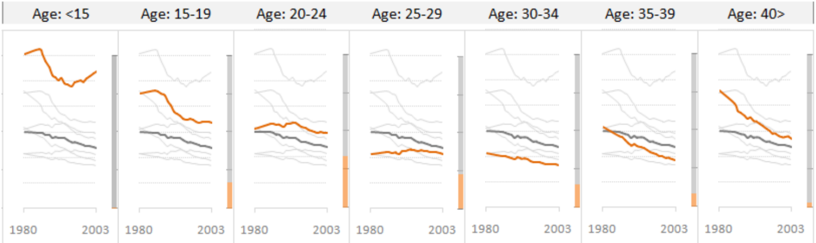
- Which group has the highest/lowest rates? When?
- Which group has an increasing/decreasing temporal trend?
- Which group has a faster/slower rate of change?

Age	1980	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003
< 15	607	624	578	523	502	492	479	498	497	519	537
$15 - 19$	451	462	449	418	379	364	347	346	337	341	337
$20 - 24$	310	328	327	328	333	326	314	301	296	298	293
$25 - 29$	213	219	216	213	224	230	224	226	221	219	211
$30 - 34$	213	203	197	189	192	189	179	176	171	171	167
$35 - 39$	317	280	265	244	241	234	219	208	200	195	186
≥ 40	461	409	374	350	339	329	309	291	283	276	268

- Which group has the highest/lowest rates? When?
- Which group has an increasing/decreasing temporal trend?
- Which group has a faster/slower rate of change?

Age	1980	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003
< 15	607	624	578	523	502	492	479	498	497	519	537
$15-19$	451	462	449	418	379	364	347	346	337	341	337
$20-24$	310	328	327	328	333	326	314	301	296	298	293
$25-29$	213	219	216	213	224	230	224	226	221	219	211
$30-34$	213	203	197	189	192	189	179	176	171	171	167
$35-39$	317	280	265	244	241	234	219	208	200	195	186
≥ 40	461	409	374	350	339	329	309	291	283	276	268

- Which group has the highest/lowest rates? When?
- Which group has an increasing/decreasing temporal trend?
- Which group has a faster/slower rate of change?



Outline

Exploratory Data Analysis (EDA)

obs	totbill	tip	sex	smoker	day	time	size
1	16.99	1.01	F	No	Sun	Night	2
2	10.34	1.66	M	No	Sun	Night	3
3	21.01	3.5	M	No	Sun	Night	3
4	23.68	3.31	M	No	Sun	Night	2
5	24.59	3.61	F	No	Sun	Night	4
6	25.29	4.71	M	No	Sun	Night	4
7	8.77	2	M	No	Sun	Night	2
8	26.88	3.12	M	No	Sun	Night	4
9	15.04	1.96	M	No	Sun	Night	2
10	14.78	3.23	M	No	Sun	Night	2
11	10.27	1.71	M	No	Sun	Night	2
12	35.26	5	F	No	Sun	Night	4
13	15.42	1.57	M	No	Sun	Night	2
14	18.43	3	M	No	Sun	Night	4
15	14.83	3.02	F	No	Sun	Night	2
16	21.58	3.92	M	No	Sun	Night	2
17	10.33	1.67	F	No	Sun	Night	3
18	16.29	3.71	M	No	Sun	Night	3
19	16.97	3.5	F	No	Sun	Night	3
20	20.65	3.35	M	No	Sat	Night	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
244	18.78	3	F	No	Thu	Night	2

Primary Analysis

- Fit a linear regression model where the tip rate as the target variable, and a single feature party size
- The fitted model is

$$\text{tip} = 0.18 - 0.01 \times \text{size}$$

which says that as the size of the dining party increases by one person (leading to a higher bill), the tip rate will decrease by 1

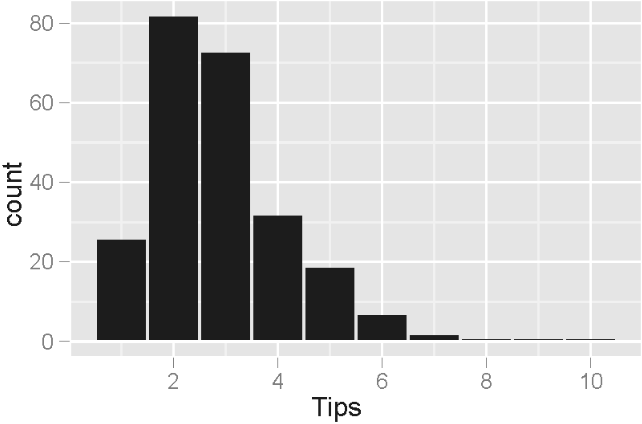


Figure: Histogram of tip amounts where the bins cover £1 increments. The distribution of values is skewed right and unimodal, as is common in distributions of small, non-negative quantities.

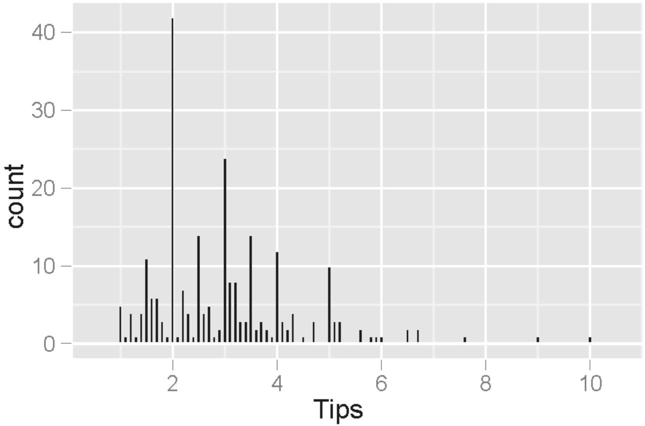


Figure: Histogram of tip amounts where the bins cover £0.10 increments. An interesting phenomenon is visible: peaks occur at the whole-dollar and half-dollar amounts, which is caused by customers picking round numbers as tips. This behaviour is common to other types of purchases too.

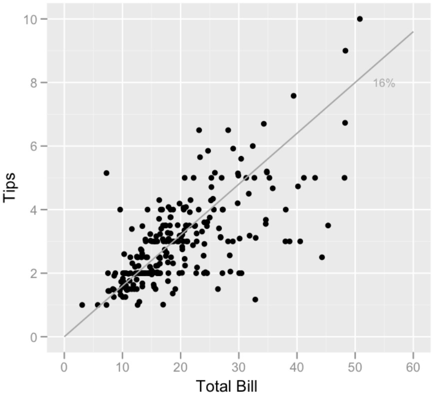


Figure: Scatterplot of tips vs. bill. Points below the line correspond to tips that are lower than expected (for that bill amount), and points above the line are higher than expected. We might expect to see a tight, positive linear association, but instead variation increases with tip amount. In particular, there are more points far away from the line in the lower right than in the upper left, indicating that more customers are very cheap than very generous.

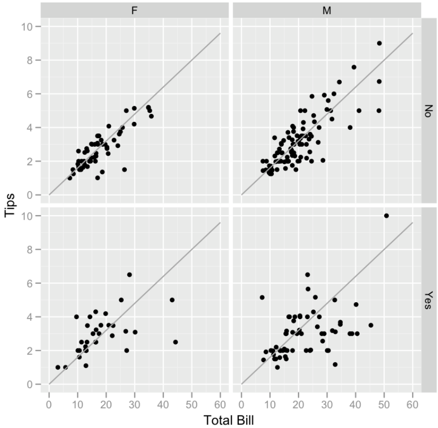
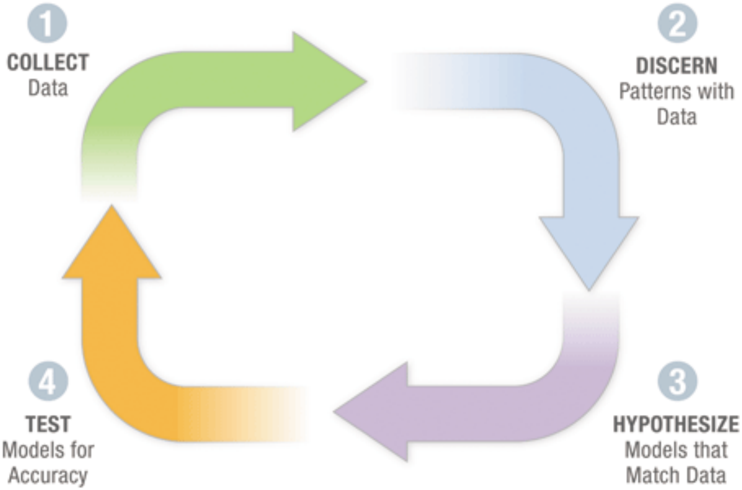


Figure: Tips vs. bill separated by gender and smoking section. Smoking parties have a lot more variability in the tips that they give. Males tend to pay the (few) higher bills, and the female non-smokers tend to be very consistent tippers (with three conspicuous exceptions shown in the sample).

EDA Outcomes

- What is learnt from the plots is different from what is illustrated by the regression model, even though the experiment was not designed to investigate any of these other trends?
- Suggests hypotheses about tipping that may not have been anticipated in advance
- Could lead to interesting follow-up experiments where the hypotheses are formally stated and tested by collecting new data

EDA Virtuous Circle



Outline

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

$\mu_x = 9$ (exact)

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

$\mu_x = 9$ (exact)
$\sigma_x^2 = 11$ (exact)

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

$\mu_x = 9$ (exact)
$\sigma_x^2 = 11$ (exact)
$\mu_y = 7.50$ (to 2 d.p.)

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

$\mu_x = 9$ (exact)
$\sigma_x^2 = 11$ (exact)
$\mu_y = 7.50$ (to 2 d.p.)
$\sigma_y^2 = 4.125 \pm .003$

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

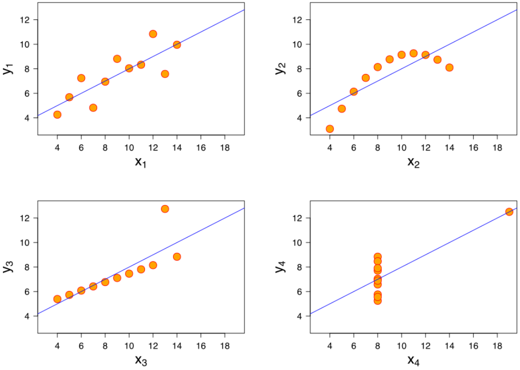
$\mu_x = 9$ (exact)
$\sigma_x^2 = 11$ (exact)
$\mu_y = 7.50$ (to 2 d.p.)
$\sigma_y^2 = 4.125 \pm .003$
$Corr(x, y) = 0.813$ (to 3 d.p.)

Anscombe's quartet (Anscombe, 1973)

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

$\mu_x = 9$ (exact)
$\sigma_x^2 = 11$ (exact)
$\mu_y = 7.50$ (to 2 d.p.)
$\sigma_y^2 = 4.125 \pm .003$
$Corr(x, y) = 0.813$ (to 3 d.p.)
Linear regression line:
$y = 3.00 + 0.500x$ (to 2 d.p.)

Anscombe's quartet (Anscombe, 1973)



$\mu_x = 9$ (exact)

$\sigma_x^2 = 11$ (exact)

$\mu_y = 7.50$ (to 2 d.p.)

$\sigma_y^2 = 4.125 \pm .003$

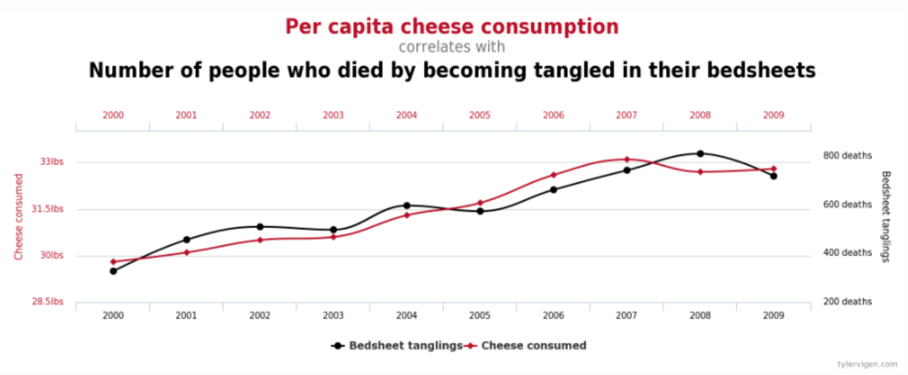
$Corr(x, y) = 0.813$ (to 3 d.p.)

Linear regression line:

$y = 3.00 + 0.500x$ (to 2 d.p.)

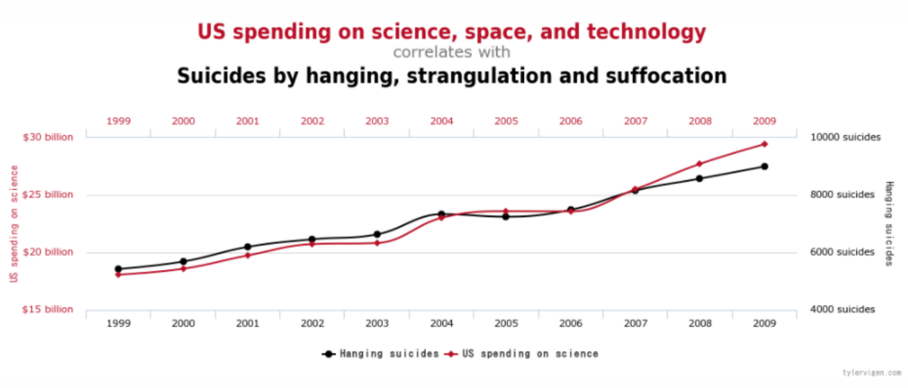
- Descriptive statistics can hide important information!

Counter Examples



http://www.tylervigen.com/spurious-correlations

Counter Examples



http://www.tylervigen.com/spurious-correlations

Outline

Time Series Visualisation (Aigner et al., 2011)

http://survey.timeviz.net/

- Data
  - Frame of Reference
    - Abstract
    - Spatial
  - Number of Variables
    - Univariate
    - Multivariate
- Time
  - Arrangement
    - Linear
    - Cyclic
  - Time Primitives
    - Instant
    - Interval
- Visualisation
  - Mapping
    - Static
    - Dynamic
  - Dimensionality
    - 2D
    - 3D


The TimeViz Browser
A Visual Survey of Visualization Techniques for Time-Oriented Data
by Christian Tominski and Wolfgang Aigner
# of Techniques: 115
Search:
How to use filters:
Want: Show me!
Indifferent: I don't care.
Hide: I'm not interested!
Data
Frame of Reference
Abstract
Spatial
Number of Variables
Univariate
Multivariate
Time
Arrangement
Linear
Cyclic
Time Primitives
Instant
Interval
Visualization
Mapping
Static
Dynamic
Dimensionality
2D
3D
Our book:
Visualization of Time-Oriented Data

Outline

Visualisation

- Language built-ins
  - R, Matlab, Octave, Mathematica ...
- OS-based
  - gnuplot
- Commerical tools
  - Tableau; Microsoft BI
- python
  - Matplotlib; Seaborn; ggplot; bokeh
- JavaScript
  - D3.js; DC.js; NVD3; Vega HighCharts/HighStock/HighMaps; plotly.js; Leaflet; MetricsGraphics.js
  - Many many others ...

Supporting Technology

- Pandas
- Flask
- Shapely
- Crossfilter
- Underscore.js
- Keen Dashboards

Python

Good news: lot of options

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Python

Bad news: lot of options!

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Python

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Python

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Python

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Python

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Python

- pandas: handy for simple plots; need to learn matplotlib to customize
- seaborn: supports some more complex visualisation approaches but still requires matplotlib knowledge to tweak. Colour schemes are a nice bonus.
- ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics (Wilkinson, 2012)
- bokeh is a robust tool if you want to set up your own visualisation server but may be overkill for the simple scenarios
- plotly generates the most interactive graphs. You can save them offline and create very rich web-based visualisations ... but commercial license

Outline

- Dataset: TalkingData Mobile User Demographics
- First pass: Pandas, matplotlib, and seaborn:

https://github.com/njtwomey/ADS/blob/master/04_data_exploration_and_visualisation/02_d3_demo/BasicVisualisations.ipynb

Flask

- A minimal Flask application:

```python
from flask import Flask
app = Flask(__name__)

@app.route('/')
def hello_world():
    return 'Hello, World!'
```

- Then run the app:

```
$ export FLASK_APP=hello.py
$ python -m flask run
```

https://github.com/njtwomey/ADS/blob/master/04_data_exploration_and_visualisation/02_d3_demo/hello.py

- Install dependencies (bower required):

```
$ npm install -g bower
$ bower install dcjs d3-queue bootstrap leaflet underscore
keen-dashboards keen-js
```

- Our real server will serve up the pandas dataframe, output as json
- The raw data: `http://localhost:5000/data`

Summary

Selected References

Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

Francis J Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1): 17–21, 1973.

Vicki Bruce, Patrick R Green, and Mark A Georgeson. *Visual perception: Physiology, psychology, & ecology*. Psychology Press, 2003.

Leland Wilkinson. The grammar of graphics. In *Handbook of Computational Statistics*, pages 375–414. Springer, 2012.