## Statement for Audio and Video Learning Resources

***Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is approximately 70-90% accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.***

***If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.***

# Introduction to Data Visualisation

Ines Arana

i.arana@rgu.ac.uk

Refs:

- E. Tufte, The Visual Display of Quantitative Information, 2nd Ed., Graphics Press LLC, 2001
- E Tufte Visual Explanations, Graphics Press LLC, 1997

# Content

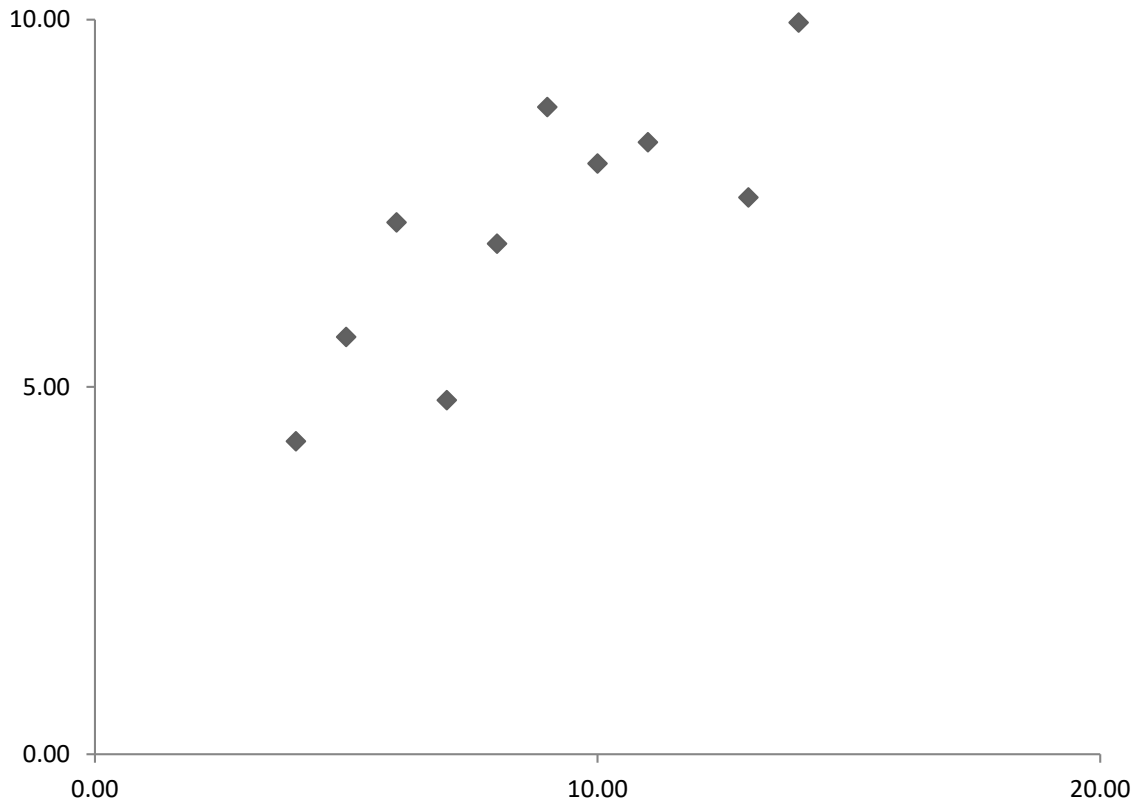
- Data
- Why visualise?
- Requirements
- Problem visualisations
- Tips for visualisation
- Summary

# Data

- How much data is there?
- In 2011 there were 1.8 zetabytes of data
- 1 zetabyte = $2^{70}$
- In 2017, 2.7 zetabytes (source: NodeGraph)
- In 2020, 44 zetabytes (https://seedscientific.com/how-much-data-is-created-every-day/)
  - Predicted **463 exabytes by 2025**
- What do we want to do with data:
  - Understand
  - Analyse
  - **Visualise**
  - Communicate

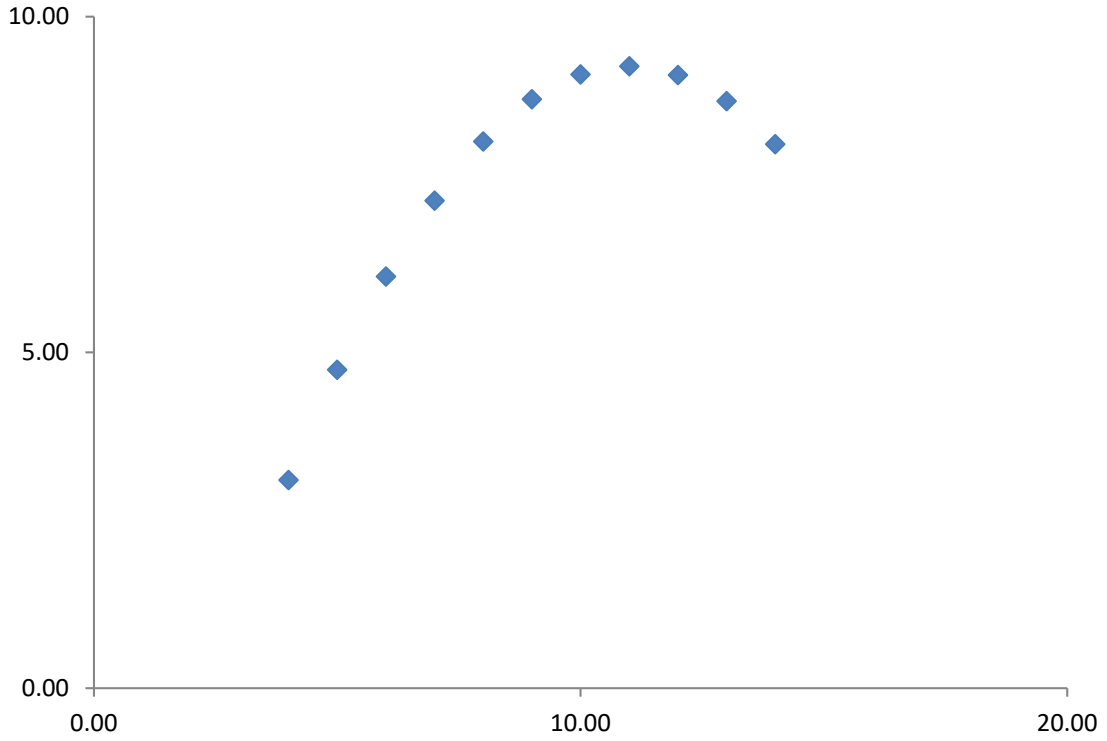
# … data

| X | Y |
|---|---|
| 10.00 | 8.04 |
| 8.00 | 6.95 |
| 13.00 | 7.58 |
| 9.00 | 8.81 |
| 11.00 | 8.33 |
| 14.00 | 9.96 |
| 6.00 | 7.24 |
| 4.00 | 4.26 |
| 12.00 | 10.84 |
| 7.00 | 4.82 |
| 5.00 | 5.68 |



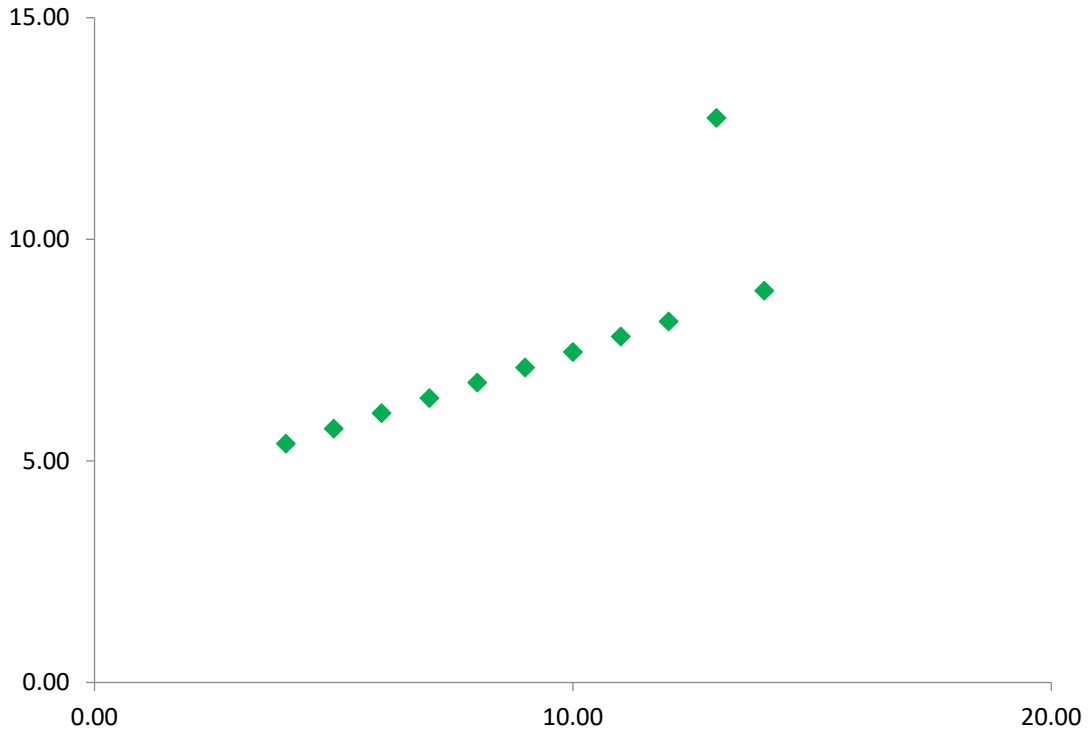
# … data

| X | Y |
|---|---|
| 10.00 | 9.14 |
| 8.00 | 8.14 |
| 13.00 | 8.74 |
| 9.00 | 8.77 |
| 11.00 | 9.26 |
| 14.00 | 8.10 |
| 6.00 | 6.13 |
| 4.00 | 3.10 |
| 12.00 | 9.13 |
| 7.00 | 7.26 |
| 5.00 | 4.74 |



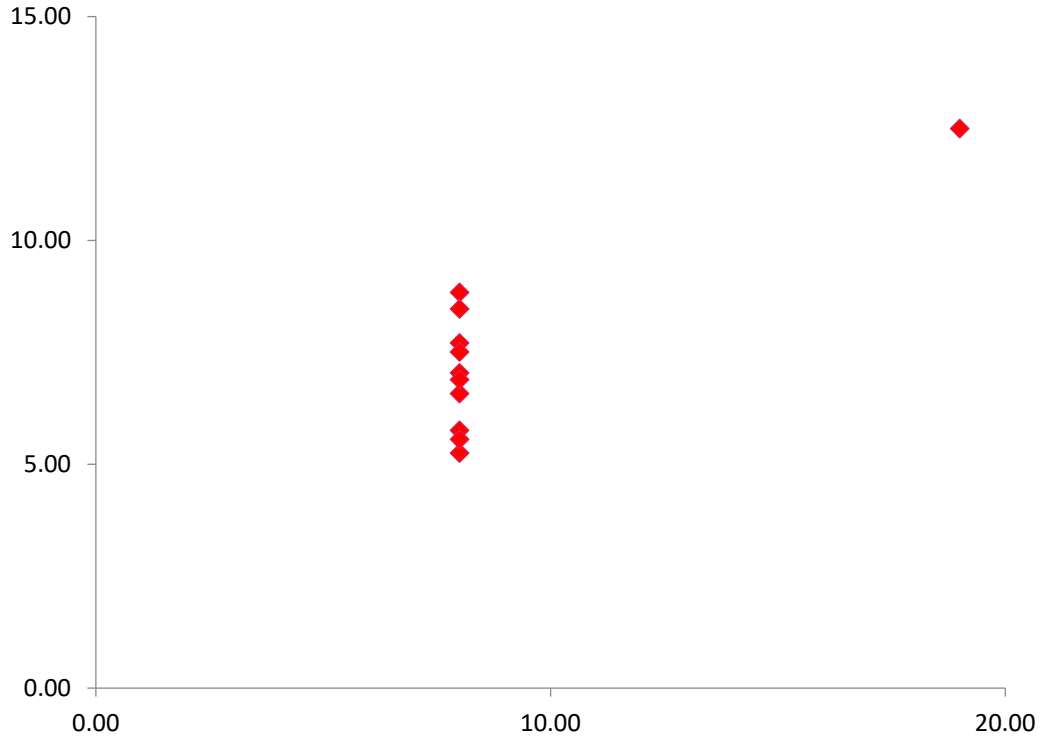
# … more data

| X | Y |
|---|---|
| 10.00 | 7.46 |
| 8.00 | 6.77 |
| 13.00 | 12.74 |
| 9.00 | 7.11 |
| 11.00 | 7.81 |
| 14.00 | 8.84 |
| 6.00 | 6.08 |
| 4.00 | 5.39 |
| 12.00 | 8.15 |
| 7.00 | 6.42 |
| 5.00 | 5.73 |



# … and more data

| X | Y |
|---|---|
| 8.00 | 6.58 |
| 8.00 | 5.76 |
| 8.00 | 7.71 |
| 8.00 | 8.84 |
| 8.00 | 8.47 |
| 8.00 | 7.04 |
| 8.00 | 5.25 |
| 19.00 | 12.50 |
| 8.00 | 5.56 |
| 8.00 | 7.51 |
| 8.00 | 6.89 |



# Data – Ascombe's Quartet

For each data set:

- X's mean: 9.0
- Y's mean: 7.5
- Equation of regression line: Y = 3 + 0.5 X
- Stdr error of estimate of slope: 0.118
- t = 4.24
- Sum of squares = 110.0
- Regression of sum of squares = 27.70
- Residual sum of squares of Y = 13:75
- Correlation coefficient = 0.82
- $r^2$= 0.67

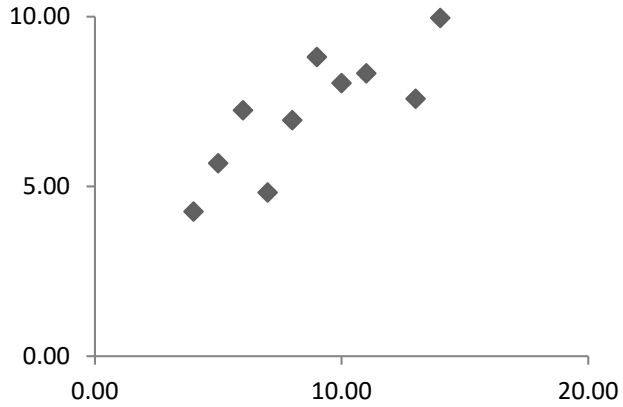
| X | Y |
|---|---|
| 10.00 | 8.04 |
| 8.00 | 6.95 |
| 13.00 | 7.58 |
| 9.00 | 8.81 |
| 11.00 | 8.33 |
| 14.00 | 9.96 |
| 6.00 | 7.24 |
| 4.00 | 4.26 |
| 12.00 | 10.84 |
| 7.00 | 4.82 |
| 5.00 | 5.68 |

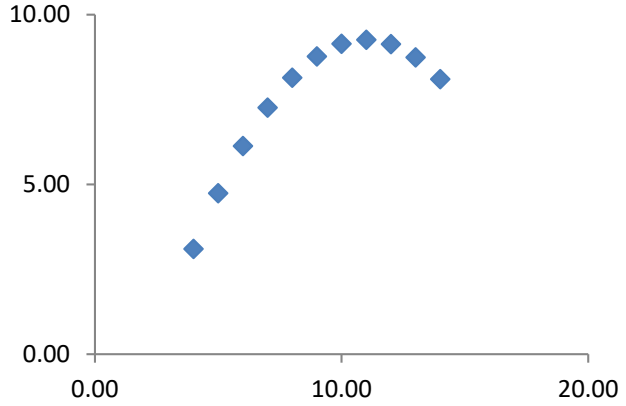
| X | Y |
|---|---|
| 10.00 | 9.14 |
| 8.00 | 8.14 |
| 13.00 | 8.74 |
| 9.00 | 8.77 |
| 11.00 | 9.26 |
| 14.00 | 8.10 |
| 6.00 | 6.13 |
| 4.00 | 3.10 |
| 12.00 | 9.13 |
| 7.00 | 7.26 |
| 5.00 | 4.74 |

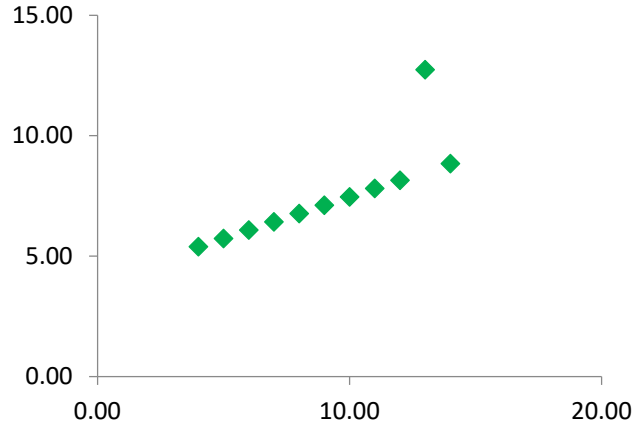
| X | Y |
|---|---|
| 10.00 | 7.46 |
| 8.00 | 6.77 |
| 13.00 | 12.74 |
| 9.00 | 7.11 |
| 11.00 | 7.81 |
| 14.00 | 8.84 |
| 6.00 | 6.08 |
| 4.00 | 5.39 |
| 12.00 | 8.15 |
| 7.00 | 6.42 |
| 5.00 | 5.73 |

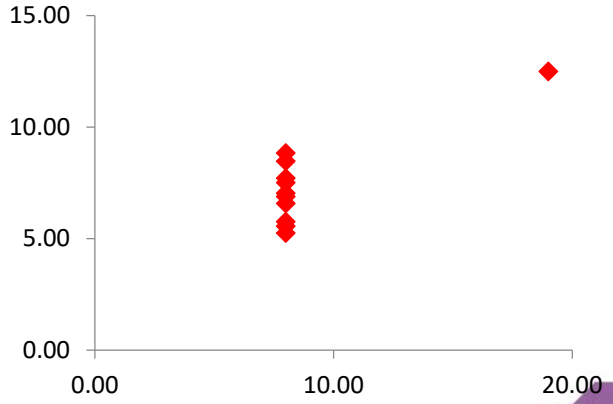
| X | Y |
|---|---|
| 8.00 | 6.58 |
| 8.00 | 5.76 |
| 8.00 | 7.71 |
| 8.00 | 8.84 |
| 8.00 | 8.47 |
| 8.00 | 7.04 |
| 8.00 | 5.25 |
| 19.00 | 12.50 |
| 8.00 | 5.56 |
| 8.00 | 7.51 |
| 8.00 | 6.89 |

# Why Visualise? Visualisation shows differences!



# Why visualise data?

- Visualisations may give us the right information to
  - Answer questions
  - Make decisions
  - Find patterns
  - Communicate a story
  - Hypothesise
  - Abstract

# Content (2)

- Data
- Why visualise?
- **Requirements**
- Problem visualisations
- Tips for visualisation
- Summary

# Visualisations - requirements

- Show the data
  - Not a pretty picture
  - Substance over presentation
- Data is informative
  - No distortion of meaning
- Lots of data presented in a compact way
- Comparison between various pieces of data should be possible
- Present data at several levels of detail depending on user
  - Summary
  - Detailed structure

# Good visualisations

- Well-designed display of data which is of interest to the viewer
- Generally
  - Multivariate  - several variables involved
  - Large amounts of data - sometimes
- Must present data in a straightforward way
  - Not misleading!

# Challenger disaster

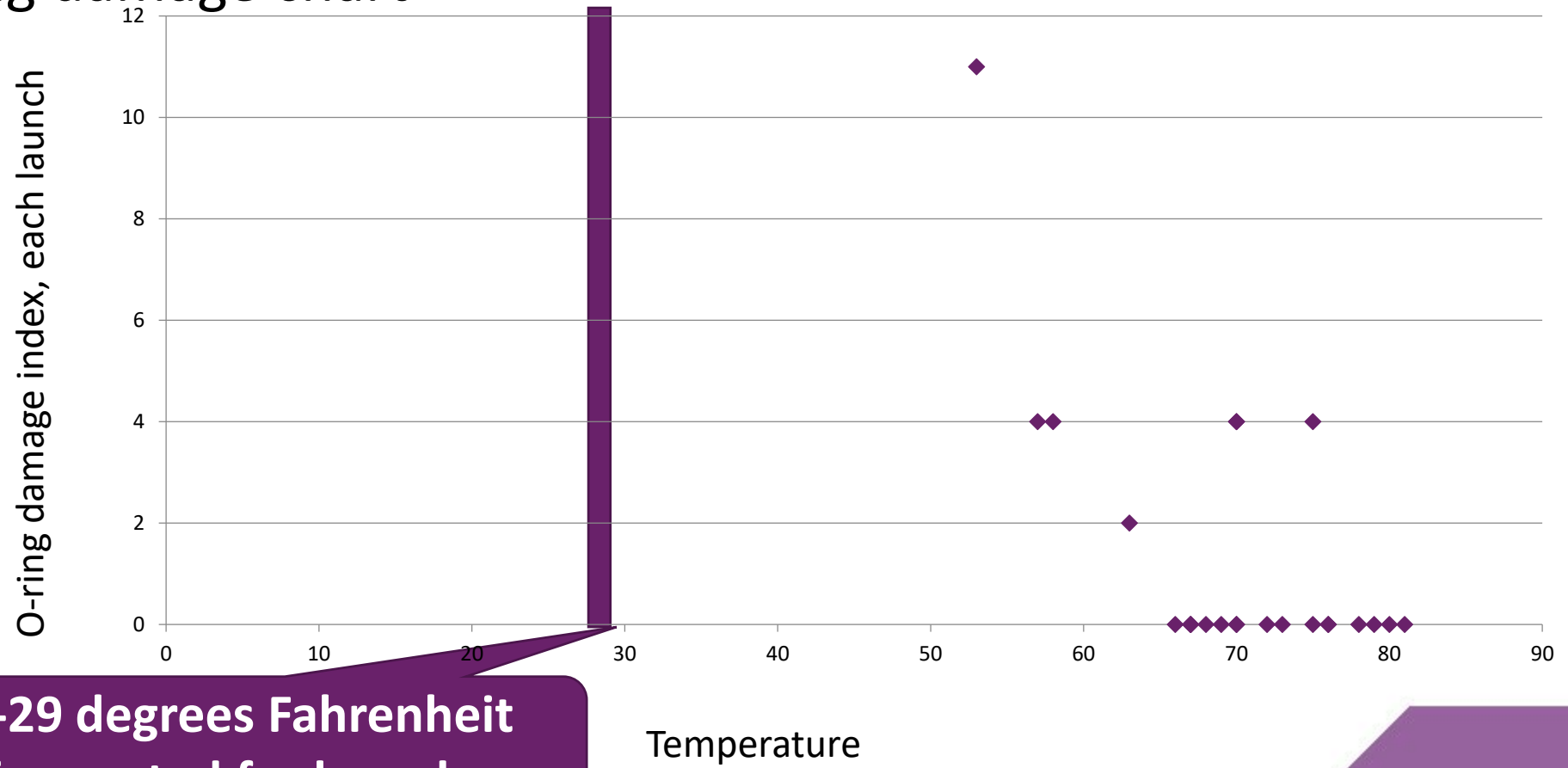
- 24h before the flight, predicted temp was 26-29°F
  - Designers opposed launch – worried about O-rings at low temperatures
  - Concerns expressed to NASA using 13 charts
  - Evidence was deemed inconclusive and launch took place
- Challenger blew up
  - Rings weakened by low temperature on the launch day:
  - Ambient  temperature was 30°F
  - Rings' temperature was 20°F

# Challenger disaster

| Temperat. | Erosion incidents | Blow by incidents | Damage index | Comments |
|---|---|---|---|---|
| 53 | 3 | 2 | 11 | *Most erosion any flight. Blow by. Back up rings heated* |
| 57 | 1 | | 4 | *Deep extensive erosion* |
| 58 | 1 | | 4 | *o-ring erosion on lauch two weeks before challenger* |
| 63 | 1 | | 2 | *o-rings showed signs of heating, but no damage* |
| 66 | | | 0 | *coolest (66) launch without o-ring problems* |
| 67 | | | 0 | |
| 67 | | | 0 | |
| 67 | | | 0 | |
| 68 | | | 0 | |
| 69 | | | 0 | |
| 70 | 1 | | 4 | *Extent of erosion not fully known* |
| 70 | | | 0 | |
| 70 | 1 | | 4 | |
| 70 | | | 0 | |
| 72 | | | 0 | |
| 73 | | | 0 | |
| 75 | | | 0 | |
| 75 | | 2 | 4 | *No erosion. Soot found behind 2 primary o-rings* |
| 76 | | | 0 | |
| 76 | | | 0 | |
| 78 | | | 0 | |
| 79 | | | 0 | |
| 80 | | | ? | *o-ring condition unknown; roket casing lost at sea* |
| 81 | | | 0 | |

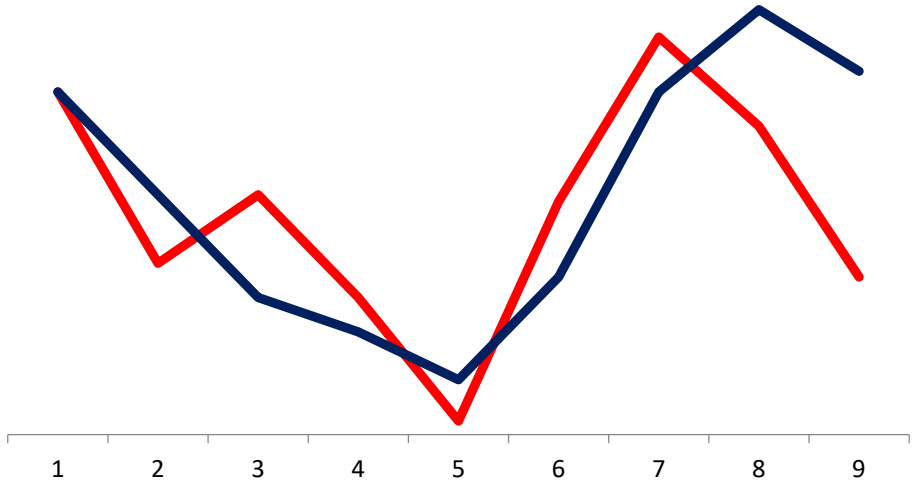
# Challenger data (Tufte,2001)

- O-ring damage chart




**26-29 degrees Fahrenheit forecasted for launch**

# Silly correlations




- Coca-Cola Stock Price (red) and average firing rates of neurons (black) from rat motor cortex over 9 days in 2004. Correlation coefficient = 0.704
- Figure adapted from: Timothy C. Marzullo, Ann Arbor, Edward G. Rantze, Cumming, Georgia, Gregory J. Gage, Ann Arbor Stock Market Behavior Predicted by Rat Neurons, Annals of Improbable Research, 22-25, July-August 2006, http://www.improbable.com/airchives/paperair/volume12/v12i4/rats-12-4.pdf [accessed 24/01/2022]

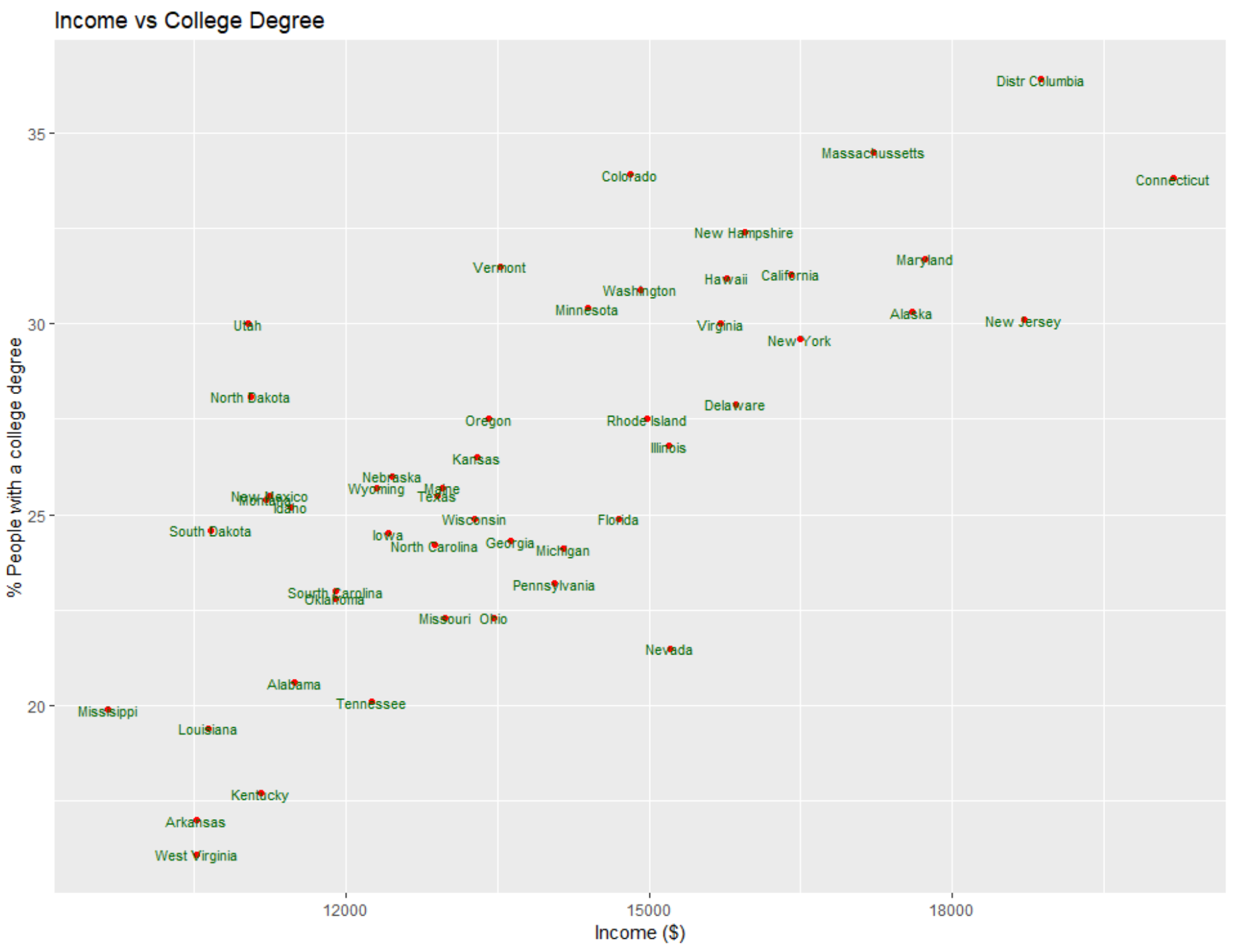
| State | % College degree | Income |
|---|---|---|
| Alabama | 20.6 | 11486 |
| Alaska | 30.3 | 17610 |
| Arkansas | 17 | 10520 |
| California | 31.3 | 16409 |
| Colorado | 33.9 | 14821 |
| Connecticut | 33.8 | 20189 |
| Delaware | 27.9 | 15854 |
| Distr Columbia | 36.4 | 18881 |
| Florida | 24.9 | 14698 |
| Georgia | 24.3 | 13631 |
| Hawaii | 31.2 | 15770 |
| Idaho | 25.2 | 11457 |
| Illinois | 26.8 | 15201 |
| Iowa | 24.5 | 12422 |
| Kansas | 26.5 | 13300 |
| Kentucky | 17.7 | 11153 |
| Louisiana | 19.4 | 10635 |
| Maine | 25.7 | 12957 |
| Maryland | 31.7 | 17730 |
| Massachussetts | 34.5 | 17224 |
| Michigan | 24.1 | 14154 |
| Minnesota | 30.4 | 14389 |
| Mississippi | 19.9 | 9648 |
| Missouri | 22.3 | 12985 |
| Montana | 25.4 | 11213 |
| Nebraska | 26 | 12452 |
| Nevada | 21.5 | 15214 |
| New Hampshire | 32.4 | 15955 |
| New Jersey | 30.1 | 18714 |
| New Mexico | 25.5 | 11246 |
| New York | 29.6 | 16501 |
| North Carolina | 24.2 | 12885 |
| North Dakota | 28.1 | 11051 |
| Ohio | 22.3 | 13461 |
| Oklahoma | 22.8 | 11893 |
| Oregon | 27.5 | 13418 |
| Pennsylvania | 23.2 | 14068 |
| Rhode Island | 27.5 | 14981 |
| Sourth Carolina | 23 | 11897 |
| South Dakota | 24.6 | 10661 |
| Tennessee | 20.1 | 12255 |
| Texas | 25.5 | 12904 |
| Utah | 30 | 11029 |
| Vermont | 31.5 | 13527 |
| Virginia | 30 | 15713 |
| Washington | 30.9 | 14923 |
| West Virginia | 16.1 | 10520 |
| Wisconsin | 24.9 | 13276 |
| Wyoming | 25.7 | 12311 |

Is there a correlation between % population with a college degree and income?

ROBERT GORDON UNIVERSITY ABERDEEN

# State income vs. college degree

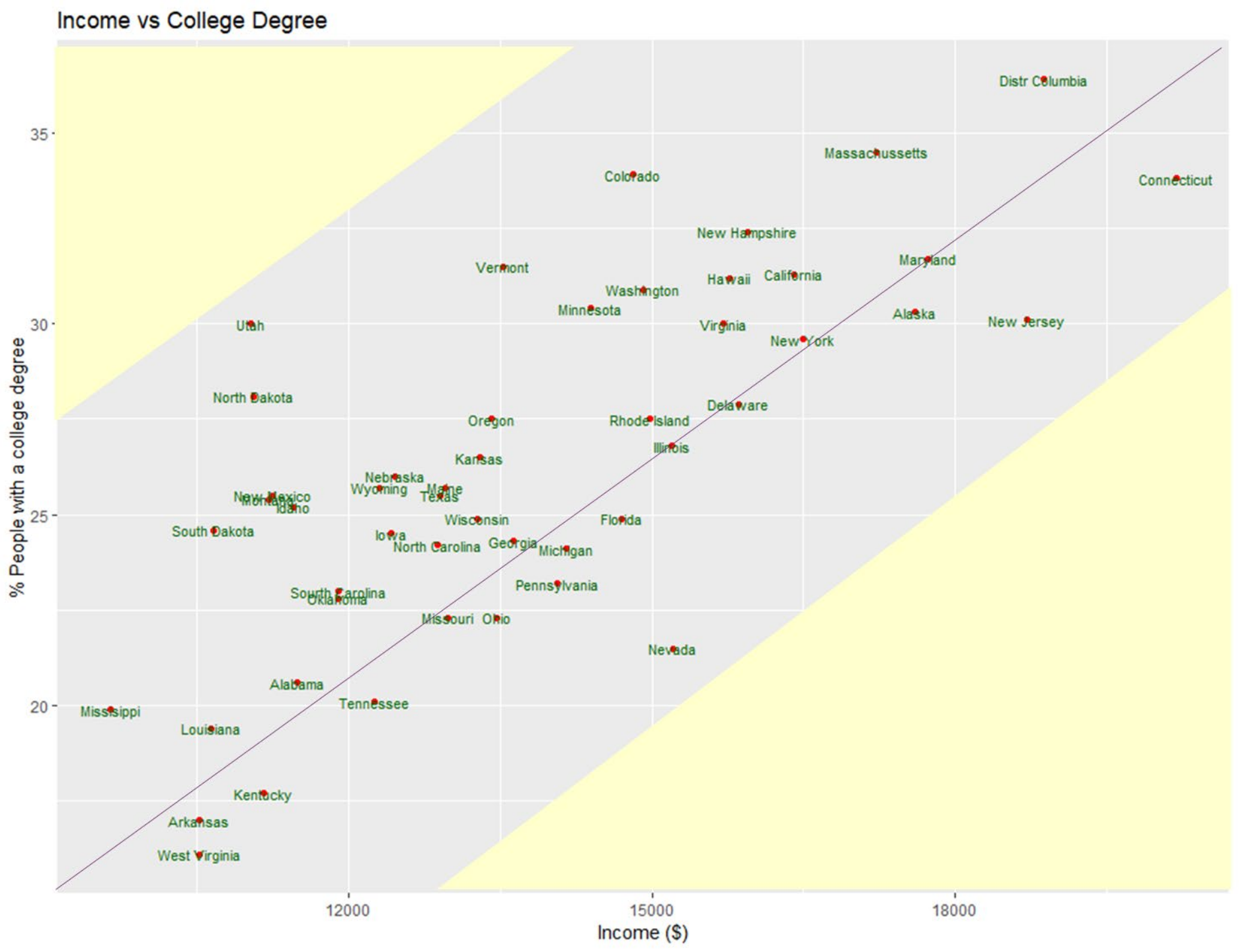
Data somehow correlated
No state with high (low) value of Income with low (high) value of % college degree



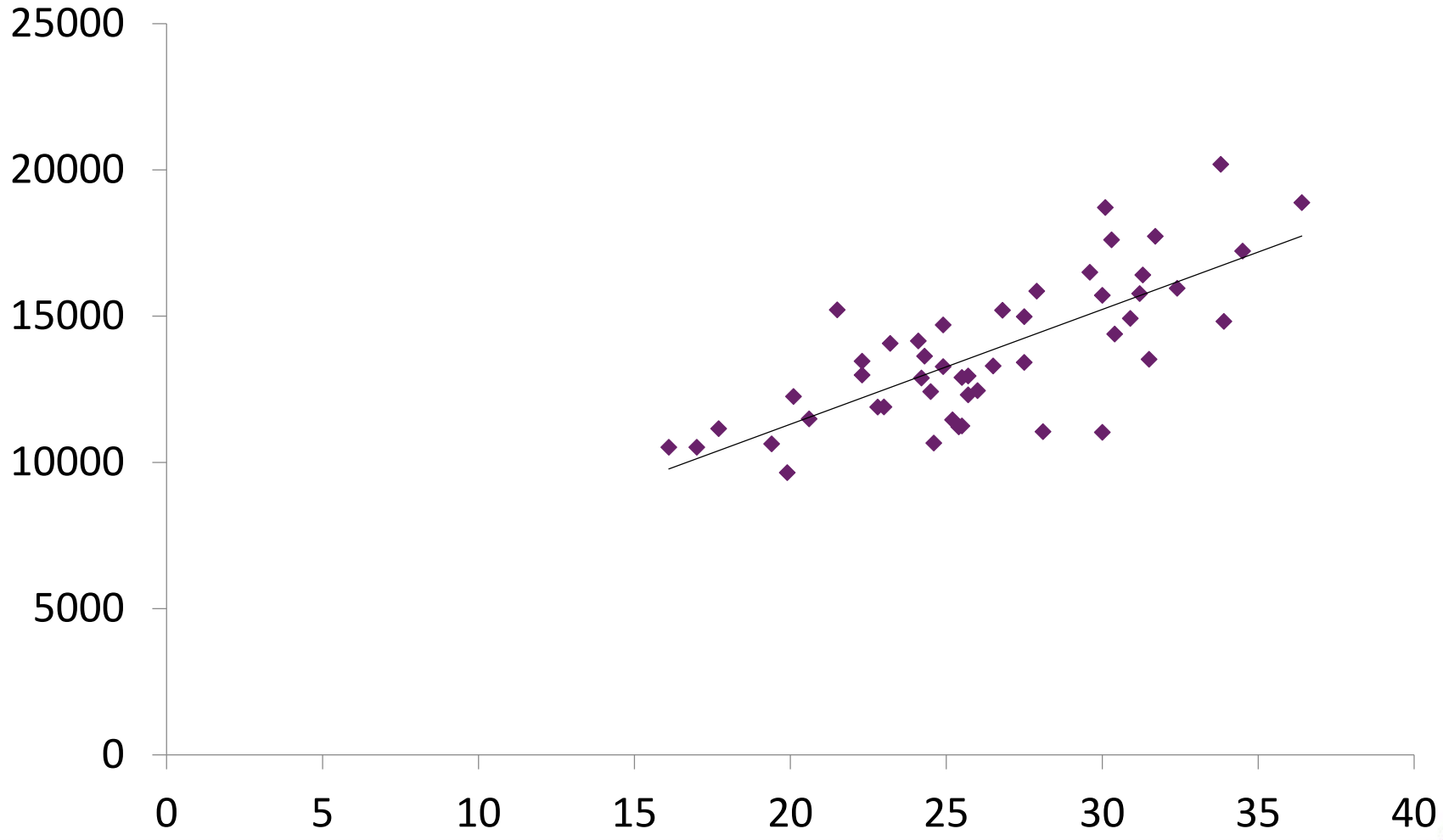
ROBERT GORDON UNIVERSITY ABERDEEN

# … state income vs. college degree

Data somehow correlated
No state with high (low) value of Income with low (high) value of % college degree



# % College degree vs. income




The labels and units of measure should be shown.
Y is income ($)
x is % college degree.

# Content (3)

- Data
- Why visualise?
- Requirements
- **Problem visualisations**
- Tips for visualisation
- Summary

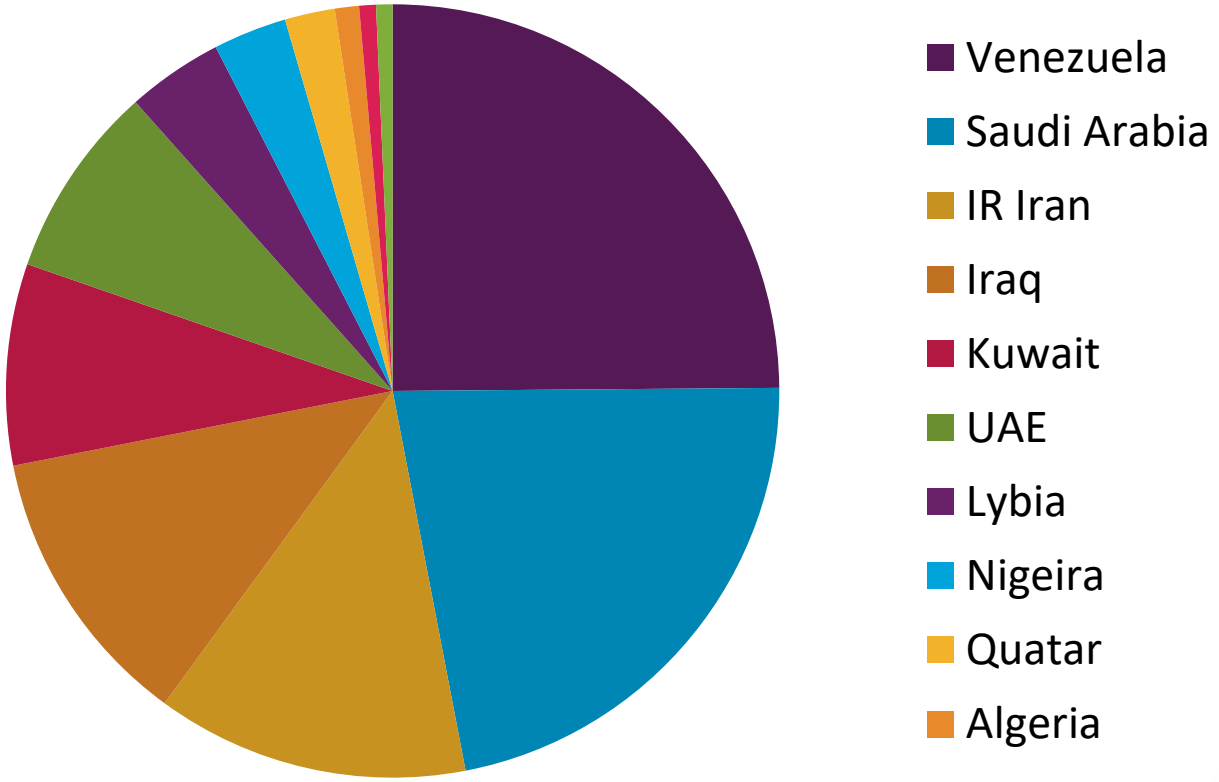
# What is wrong with the following?

- See Oil Reserves Data  in OPEC website at
  http://www.opec.org/opec_web/en/data_graphs/330.htm
  [accessed 24/1/2022]

- See the 2 data charts in the next 2 slides

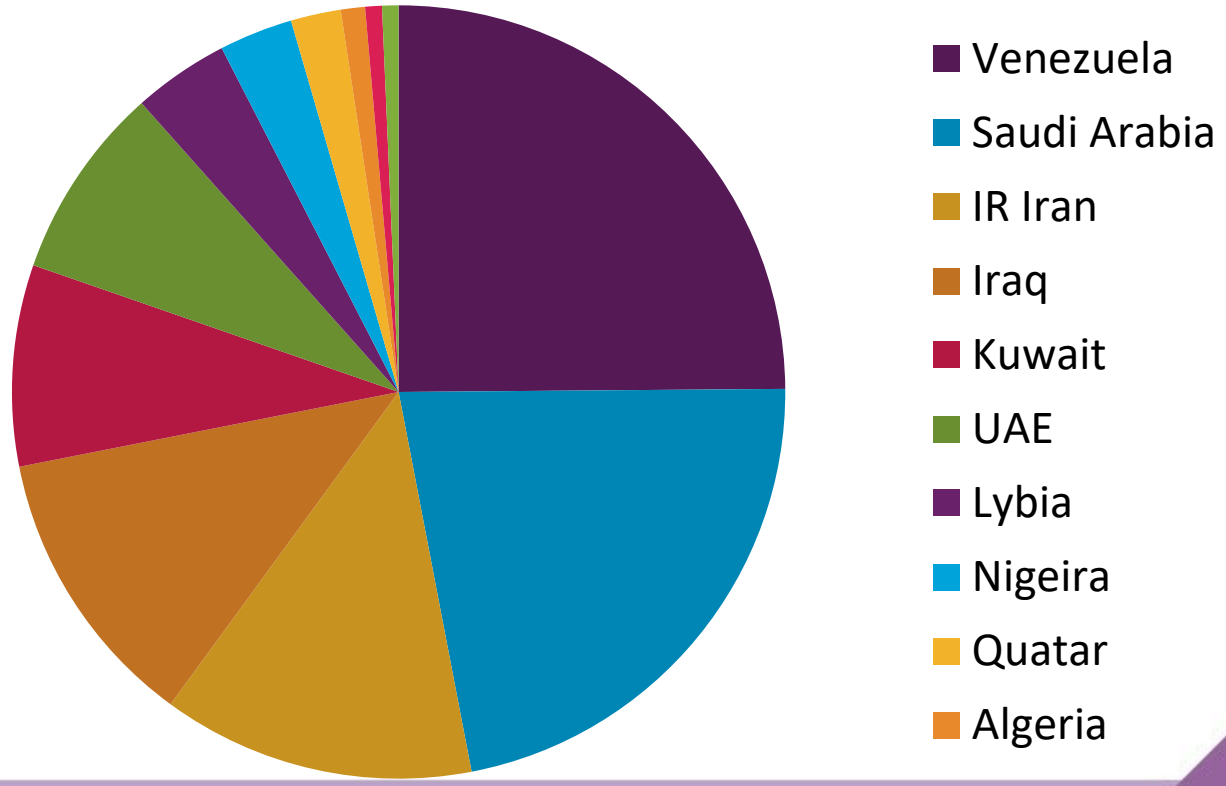
# Pie chart – Proven Oil Reserves 2014

- Proven crude oil reserves



# … pie chart

- What is Lybia's share?
- Is Algeria's share bigger or smaller than Lybia's

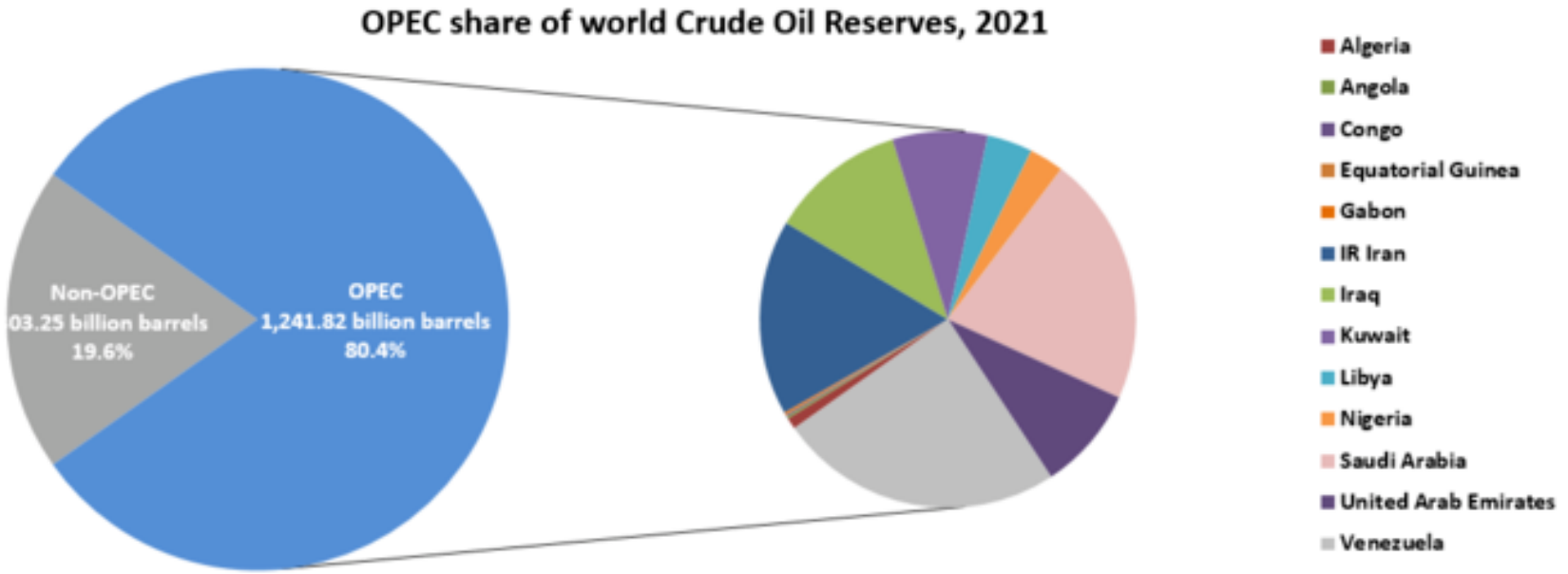


# … pie chart (cont)

- Pie shows percentages, not actual figures
- Data used is below, measured in billion barrels (OPEC share)

| | |
|---|---|
| Venezuela | 299.95 |
| Saudi Arabia | 266.58 |
| IR Iran | 157.53 |
| Iraq | 143.07 |
| Kuwait | 101.5 |
| UAE | 97.8 |
| Lybia | 48.36 |
| Nigeira | 37.07 |
| Quatar | 25.24 |
| Algeria | 12.2 |
| Angola | 8.42 |
| Ecuador | 8.27 |

# New data – OPEC share of world crude reserves 2021




This time there is a table below the plot with actual amounts and percentages

OPEC proven crude oil reserves , at end 2021 (billion barrels, OPEC share)

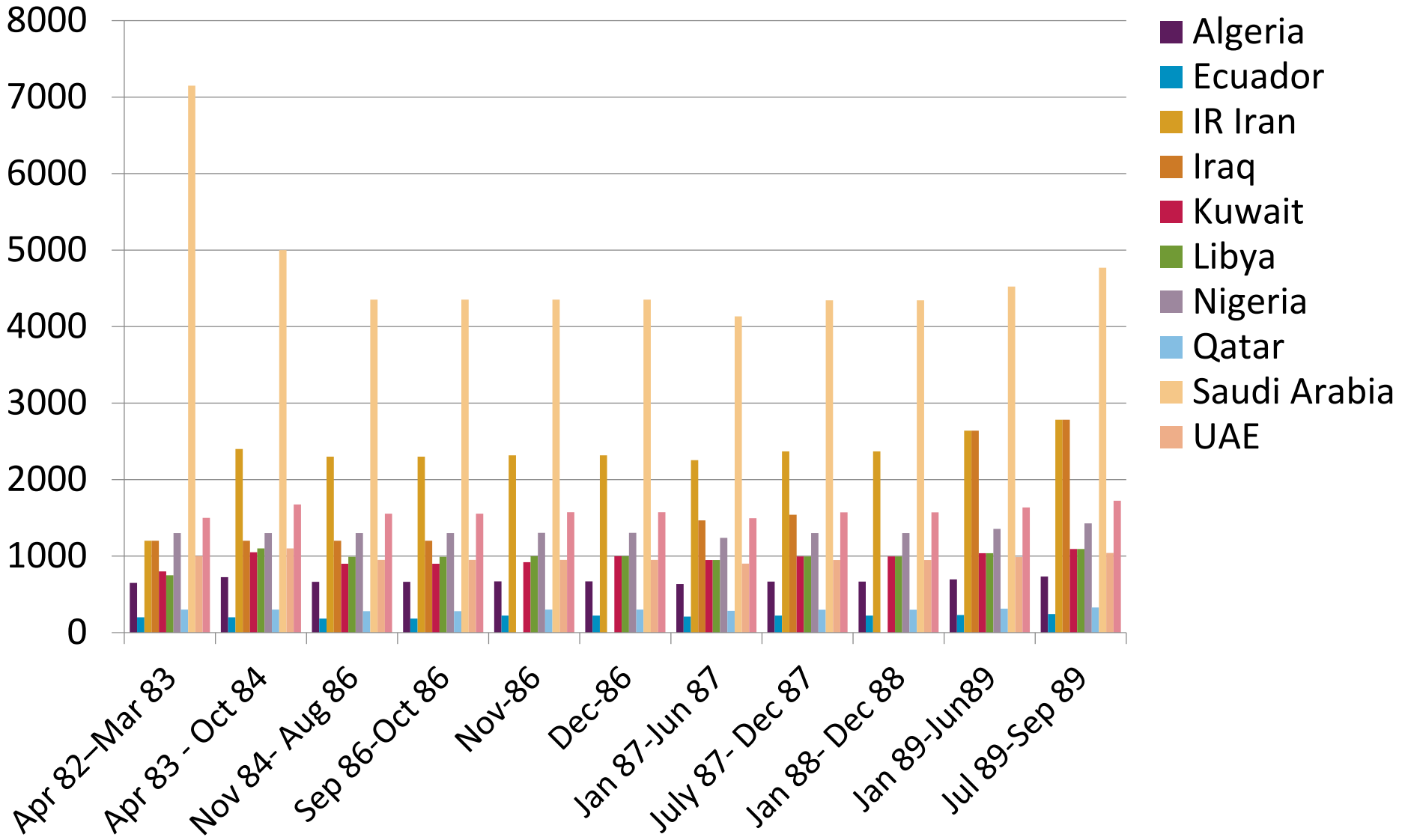
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Venezuela | 303.47 | 24.4% | United Arab Emirates | 111.00 | 8.9% | Algeria | 12.20 | 1.0% | Equatorial Guinea | 1.10 | 0.1% |
| Saudi Arabia | 267.19 | 21.5% | Kuwait | 101.50 | 8.2% | Angola | 2.52 | 0.2% | | | |
| IR Iran | 208.60 | 16.8% | Libya | 48.36 | 3.9% | Gabon | 2.00 | 0.2% | | | |
| Iraq | 145.02 | 11.7% | Nigeria | 37.05 | 3.0% | Congo | 1.81 | 0.1% | | | |

ROBERT GORDON UNIVERSITY ABERDEEN

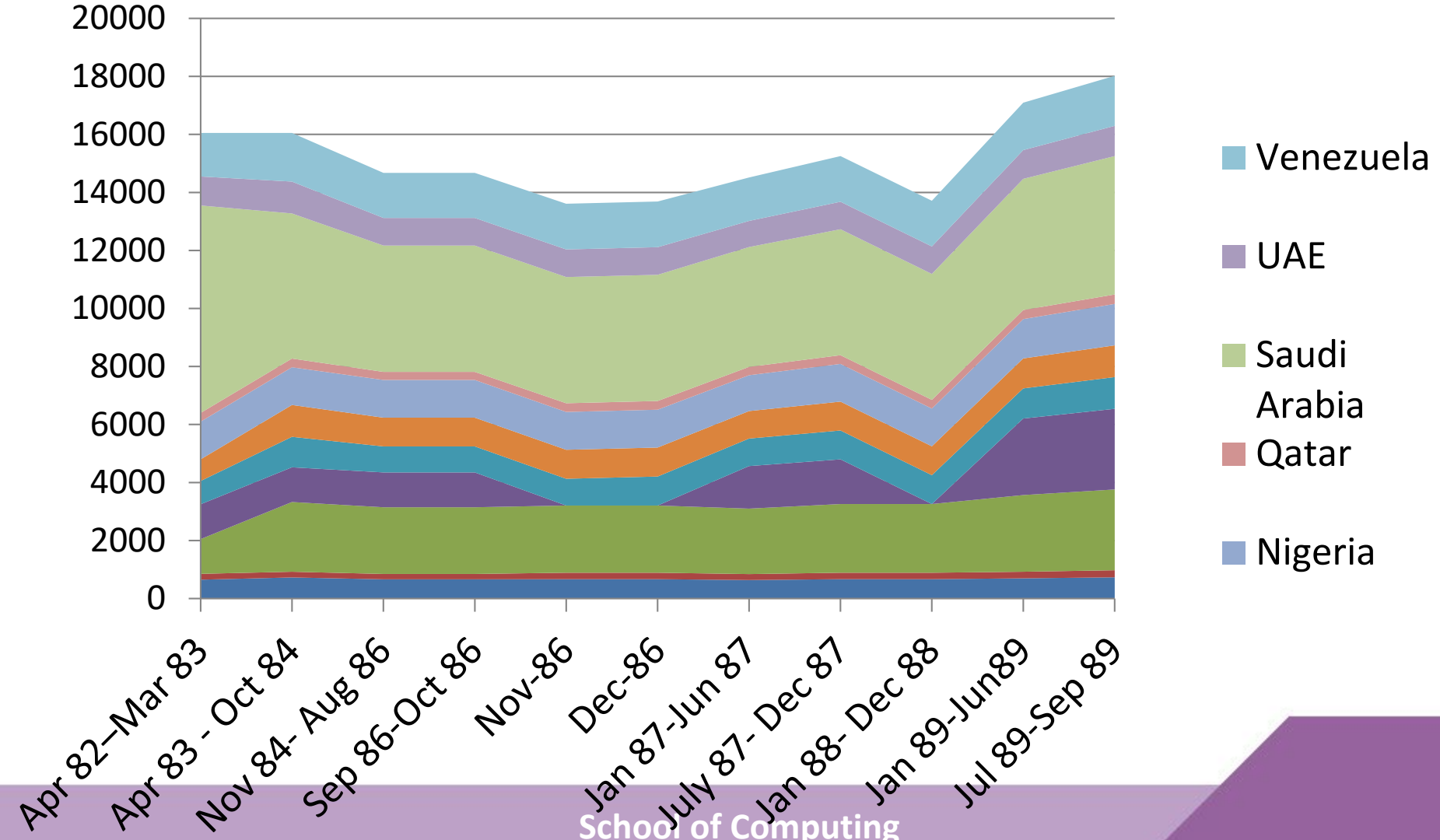
# Crude Oil Production Allocations (1000 b/d)

| | Apr 82– Mar 83 | Apr 83 - Oct 84 | Nov 84- Aug 86 | Sep 86- Oct 86 | Nov-86 | Dec-86 | Jan 87-Jun 87 | July 87- Dec 87 | Jan 88- Dec 88 | Jan 89- Jun89 | Jul 89- Sep 89 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algeria** | 650 | 725 | 663 | 663 | 669 | 669 | 635 | 667 | 667 | 695 | 733 |
| **Ecuador** | 200 | 200 | 183 | 183 | 221 | 221 | 210 | 221 | 221 | 230 | 242 |
| **IR Iran** | 1,200 | 2,400 | 2,300 | 2,300 | 2,317 | 2,317 | 2,255 | 2,369 | 2,369 | 2,640 | 2,783 |
| **Iraq** | 1,200 | 1,200 | 1,200 | 1,200 | – | – | 1,466 | 1,540 | – | 2,640 | 2,783 |
| **Kuwait** | 800 | 1,050 | 900 | 900 | 921 | 999 | 948 | 996 | 996 | 1,037 | 1,093 |
| **Libya** | 750 | 1,100 | 990 | 990 | 999 | 999 | 948 | 996 | 996 | 1,037 | 1,093 |
| **Nigeria** | 1,300 | 1,300 | 1,300 | 1,300 | 1,304 | 1,304 | 1,238 | 1,301 | 1,301 | 1,355 | 1,428 |
| **Qatar** | 300 | 300 | 280 | 280 | 300 | 300 | 285 | 299 | 299 | 312 | 329 |
| **Saudi Arabia** | 7,150 | 5,000 | 4,353 | 4,353 | 4,353 | 4,353 | 4,133 | 4,343 | 4,343 | 4,524 | 4,769 |
| **UAE** | 1,000 | 1,100 | 950 | 950 | 950 | 950 | 902 | 948 | 948 | 988 | 1,041 |
| **Venezuela** | 1,500 | 1,675 | 1,555 | 1,555 | 1,574 | 1,574 | 1,495 | 1,571 | 1,571 | 1,636 | 1,724 |

# … crude Oil Production Allocations (1000 b/d)



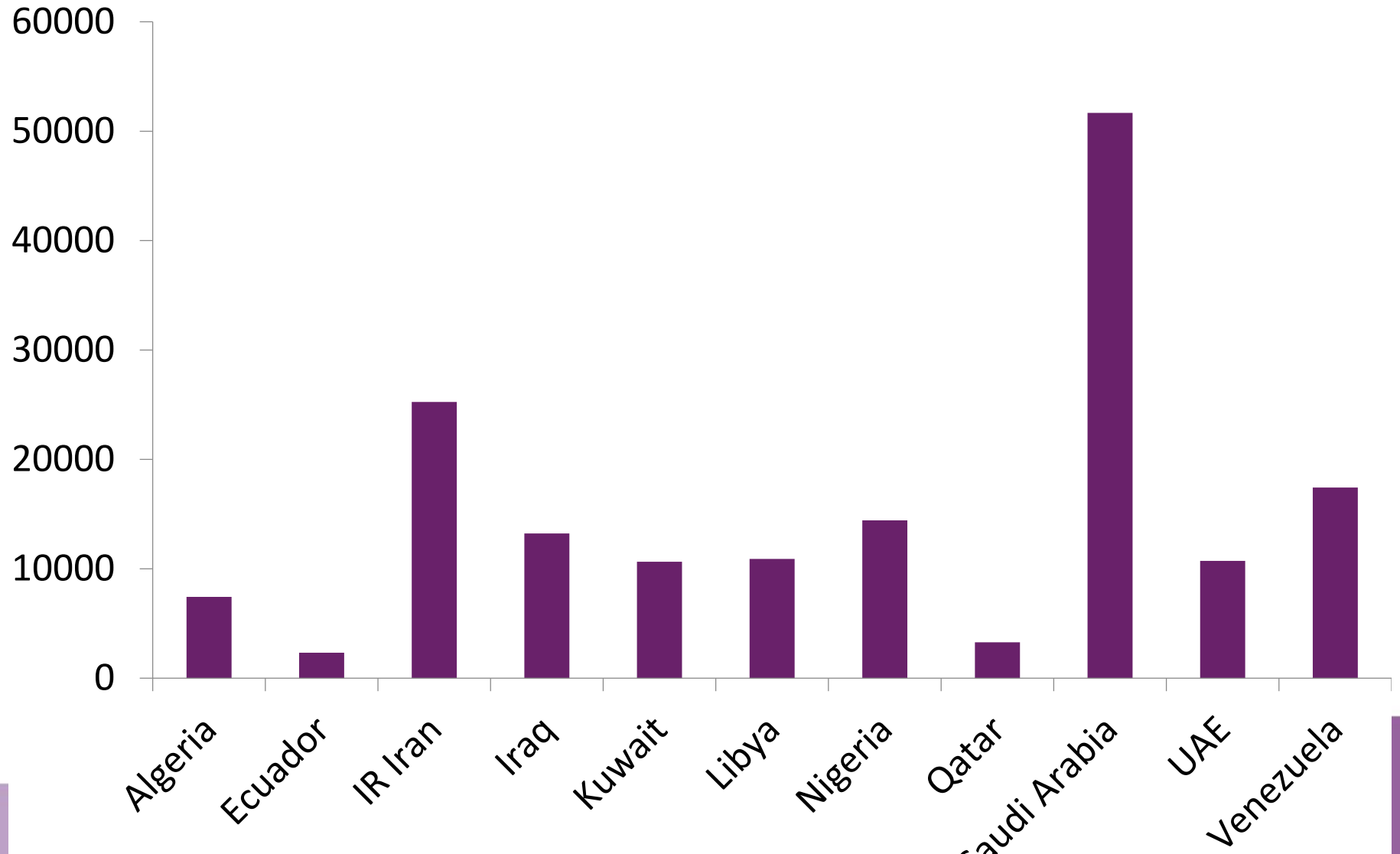
# … crude Oil Production Allocations (1000 b/d) (cont.)



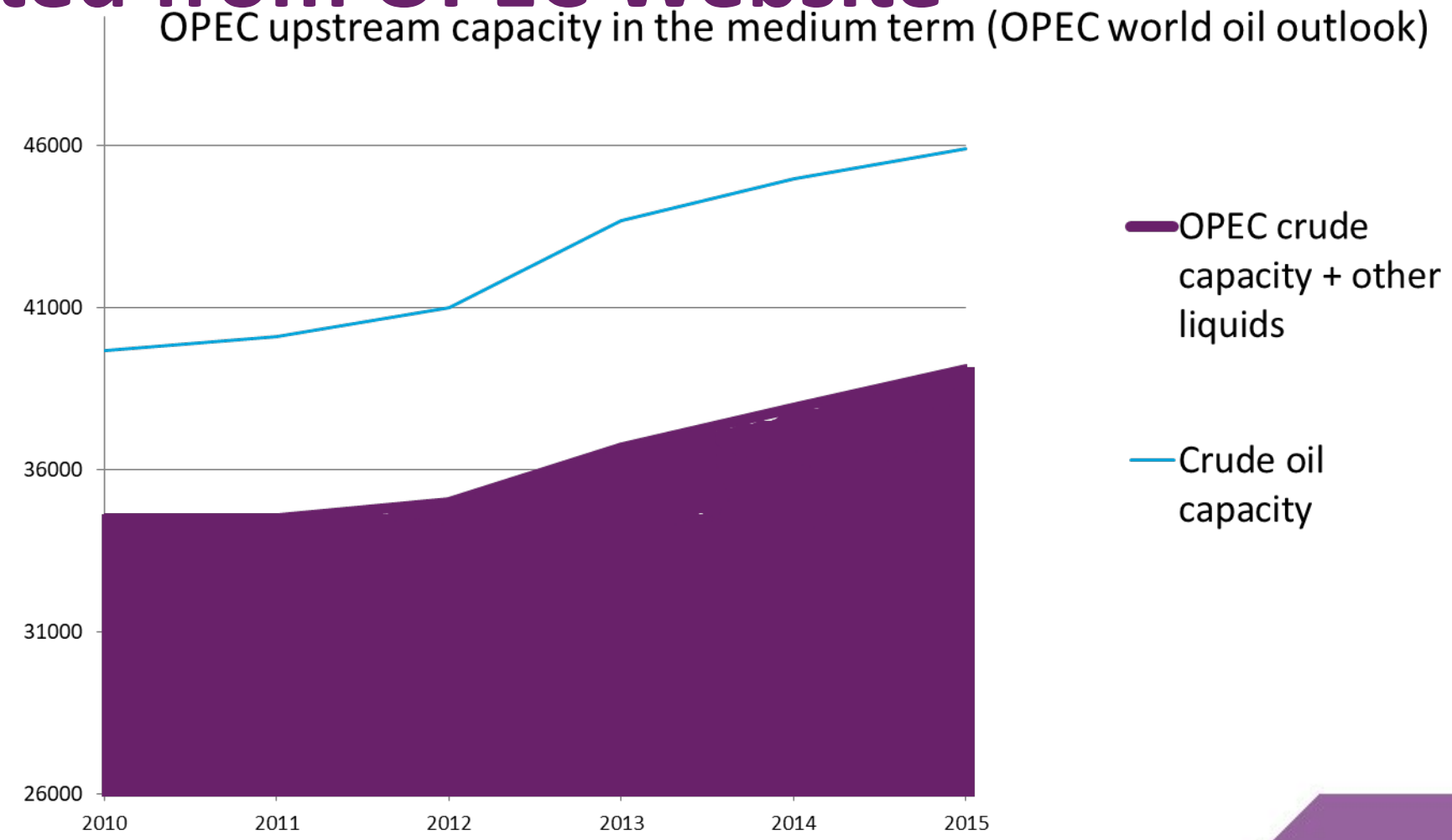
# Problems

- Time periods are varied
  - So it is difficult to decide if a country's crude production allocation has increased or decreased.
  - Barchart is too crowded
  - Other chart – difficult to assess individual country's trends as they are added on top of "previous" country's data.

# Better? - Crude production allocation Apr 82 – Sep 89

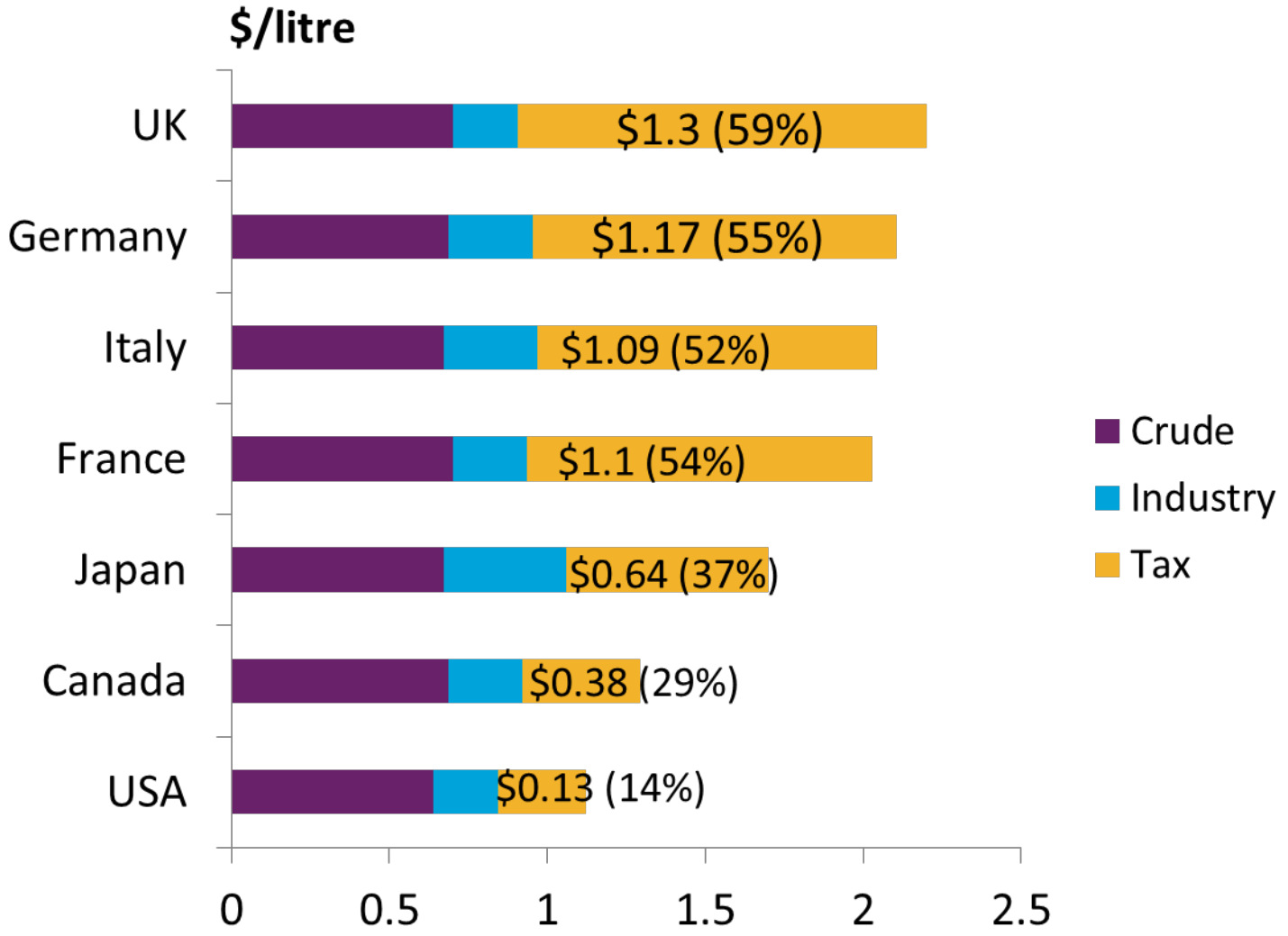


# Adapted from OPEC Website



# OPEC Data

- Adapted from figure OPEC website



# Previous 2 visualisations

- Are they clear?
- Is it easy to see who gets what from a litre of oil in the 2nd visualisation.
- Which country has a higher tax, Germany or France?
- It is easy to see that USA gets cheaper crude.

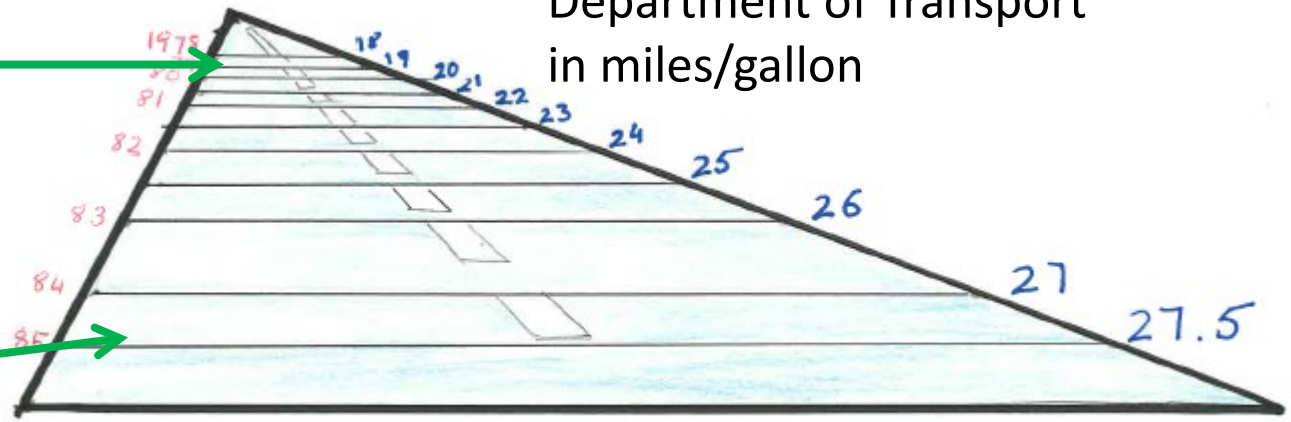
# Lies?

**Fuel economy Standards for Cars**

Set by congress and supplemented by the Department of Transport in miles/gallon

18 miles/gallon in 1978. Line is 0.6 in long

27.5 miles/gallon in 1985. Line is 5.3 in long




Adapted from
http://www.infovis-wiki.net/index.php?title=File:Lie_factor_example1_image.jpg
[accessed 29/01/2019]

# Lie Factor

- $Lie\ Factor = \frac{size\ of\ factor\ shown\ in\ graphic}{size\ of\ effect\ in\ data}$

- Good representation of data
  - Lie factor = 1
- Significant distortion
  - Lie factor < 0.95
  - Or
  - Lie factor > 1.05 ***this is what usually happens if visualisation is misleading ***

# … lie factor

- Congress and department of transportation set a series of fuel economy standards to be met by car manufacturers
  - 18 miles/gallon in 1978
  - 27.5 miles/gallon in 1985
- Increase
- $\frac{27.5 - 18.0}{18.0} \times 100 = 53\%$

# Lie factor calculations

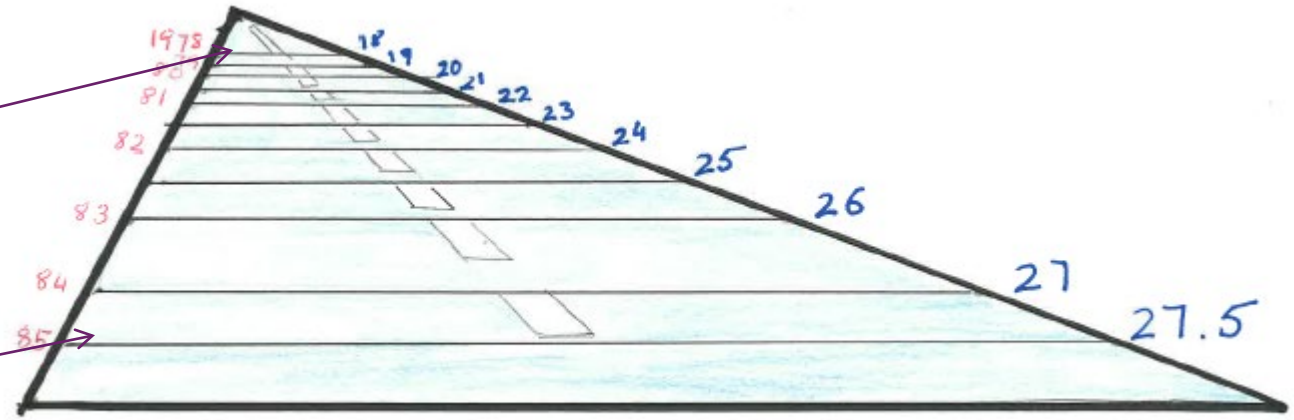
- Magnitude of change given length of lines in graph
- $\frac{5.3-0.6}{0.6} \times 100 = 783\%$
- Lie factor $= \frac{783}{53}$ = 14.8

**Fuel economy Standards for Cars**

Set by congress and supplemented by the

18 miles/gallon in 1978. Line is 0.6 in long

27.5 miles/gallon in 1985. Line is 5.3 in long



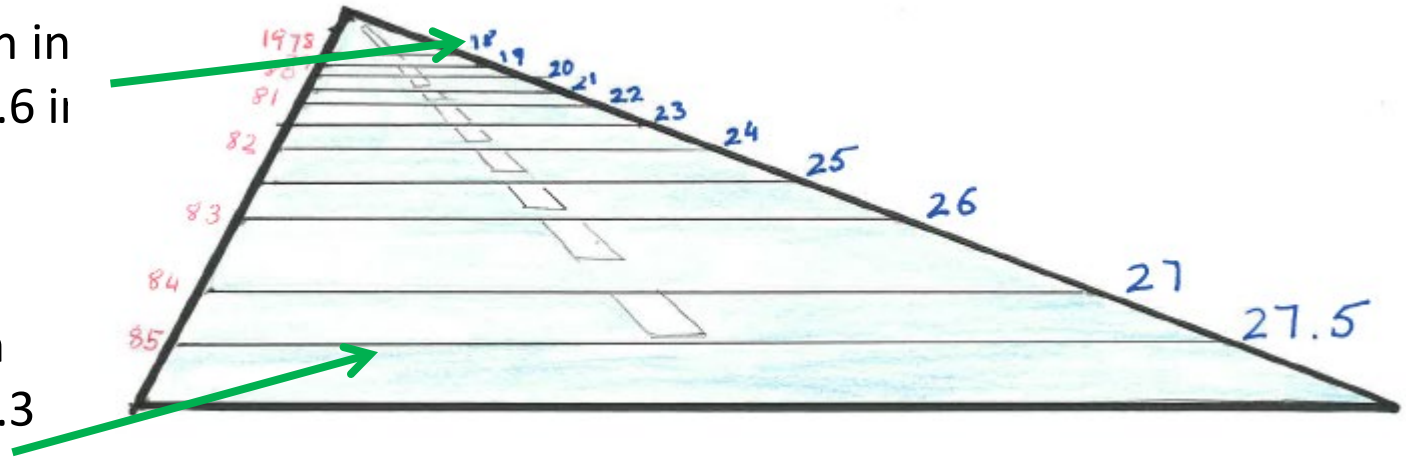
# Lie factor example - observations

- Things to note
  - Here present is closest to viewer – unusual on roads!
  - Date fontsize is constant, fuel economy fontsize is not

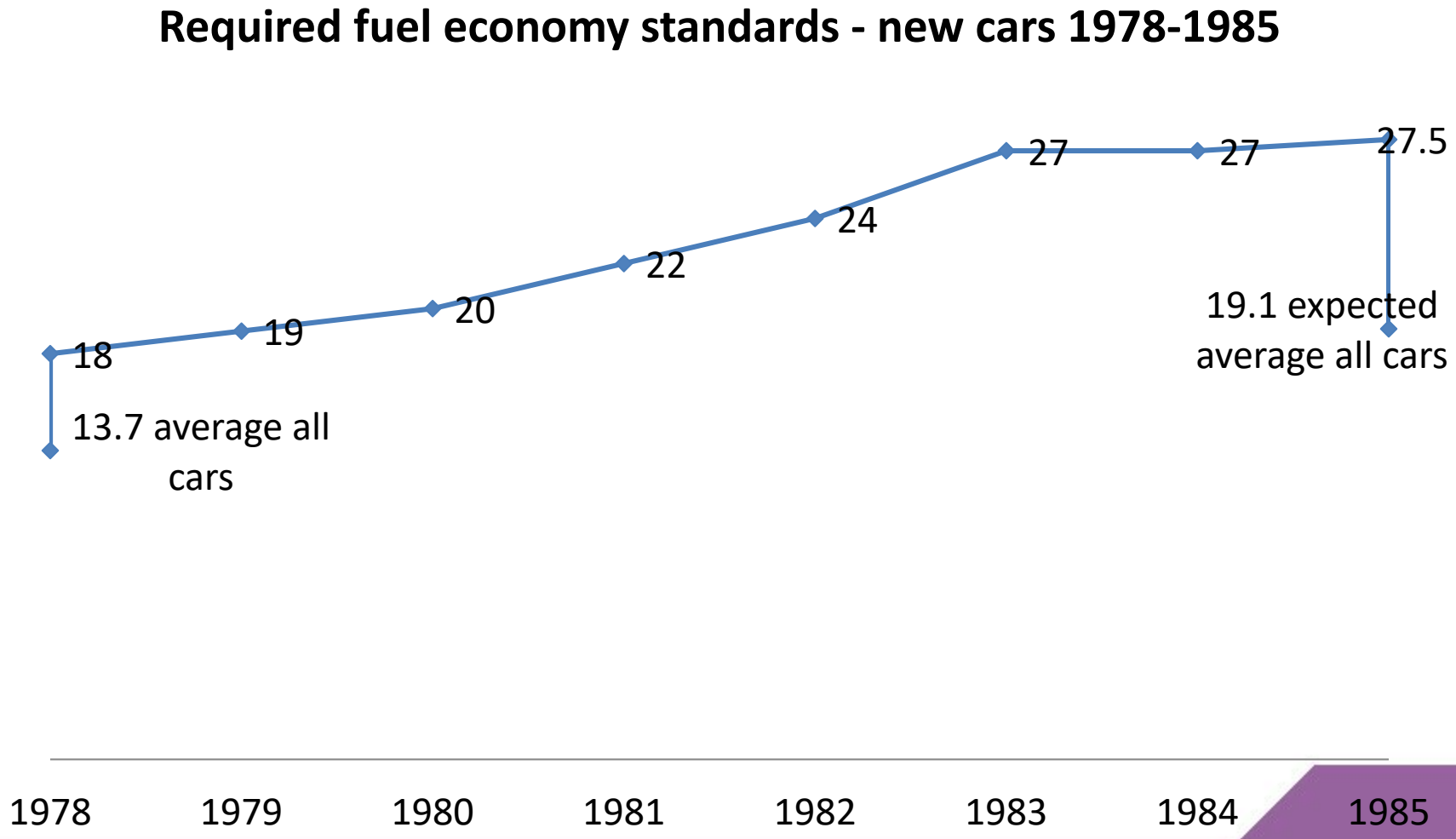
**Fuel economy Standards for Cars**
Set by congress and supplemented by the

18 miles/gallon in 1978. Line is 0.6 ii long

27.5 miles/gallon in 1985. Line is 5.3 in long

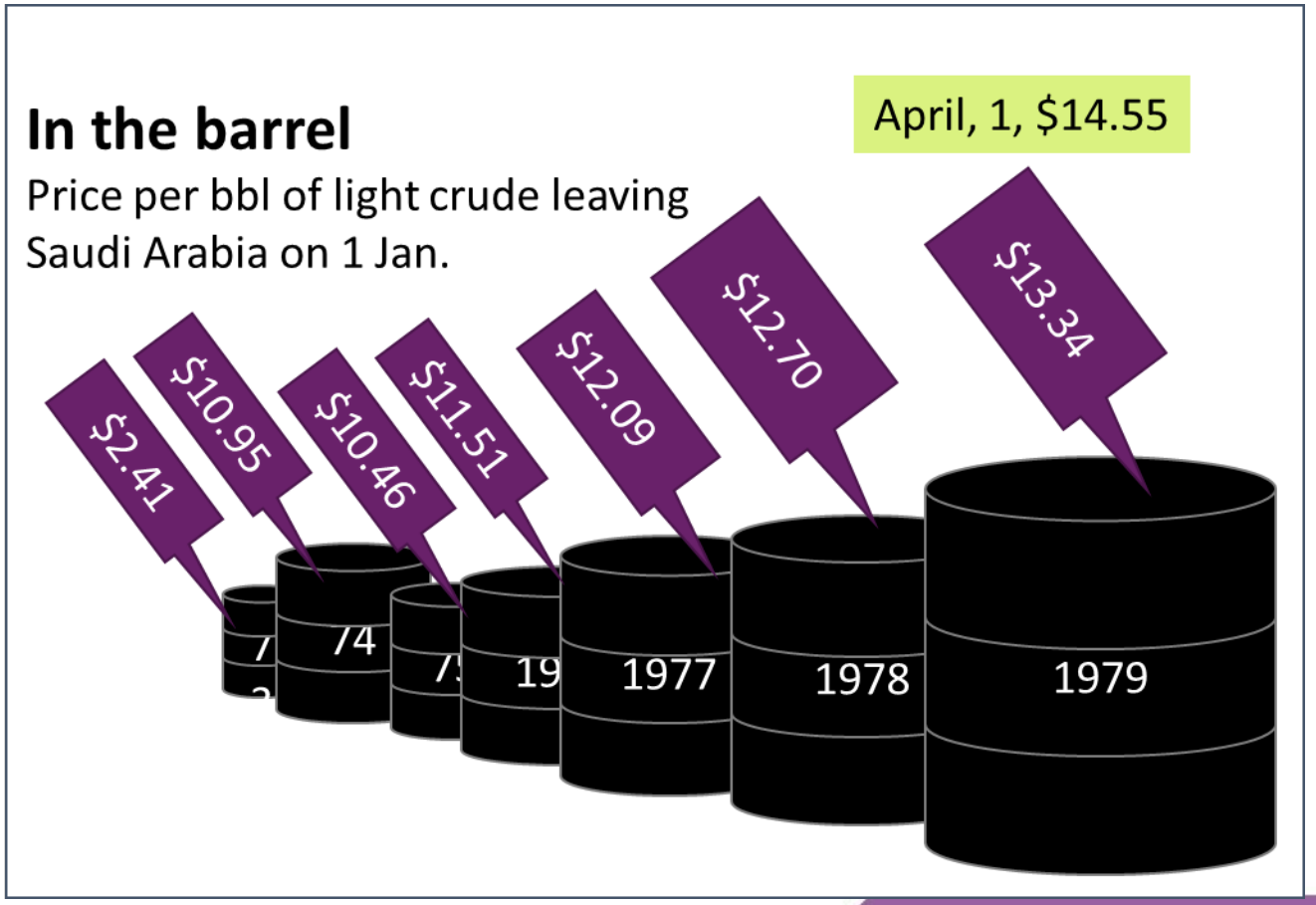


# Better representation



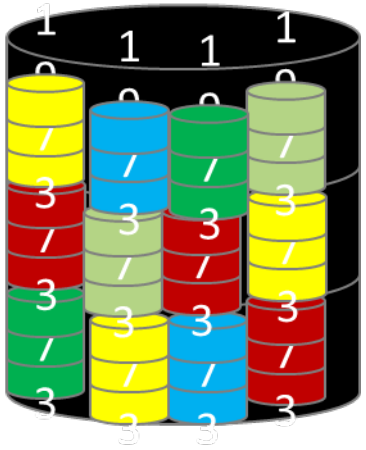
# Show data variation NOT design variation
## (adapted from Tufte 2001)

- Barrel size is misleading
- Increase of 553% in price
- BUT much bigger increase in barrel size
- Huge lie factor!

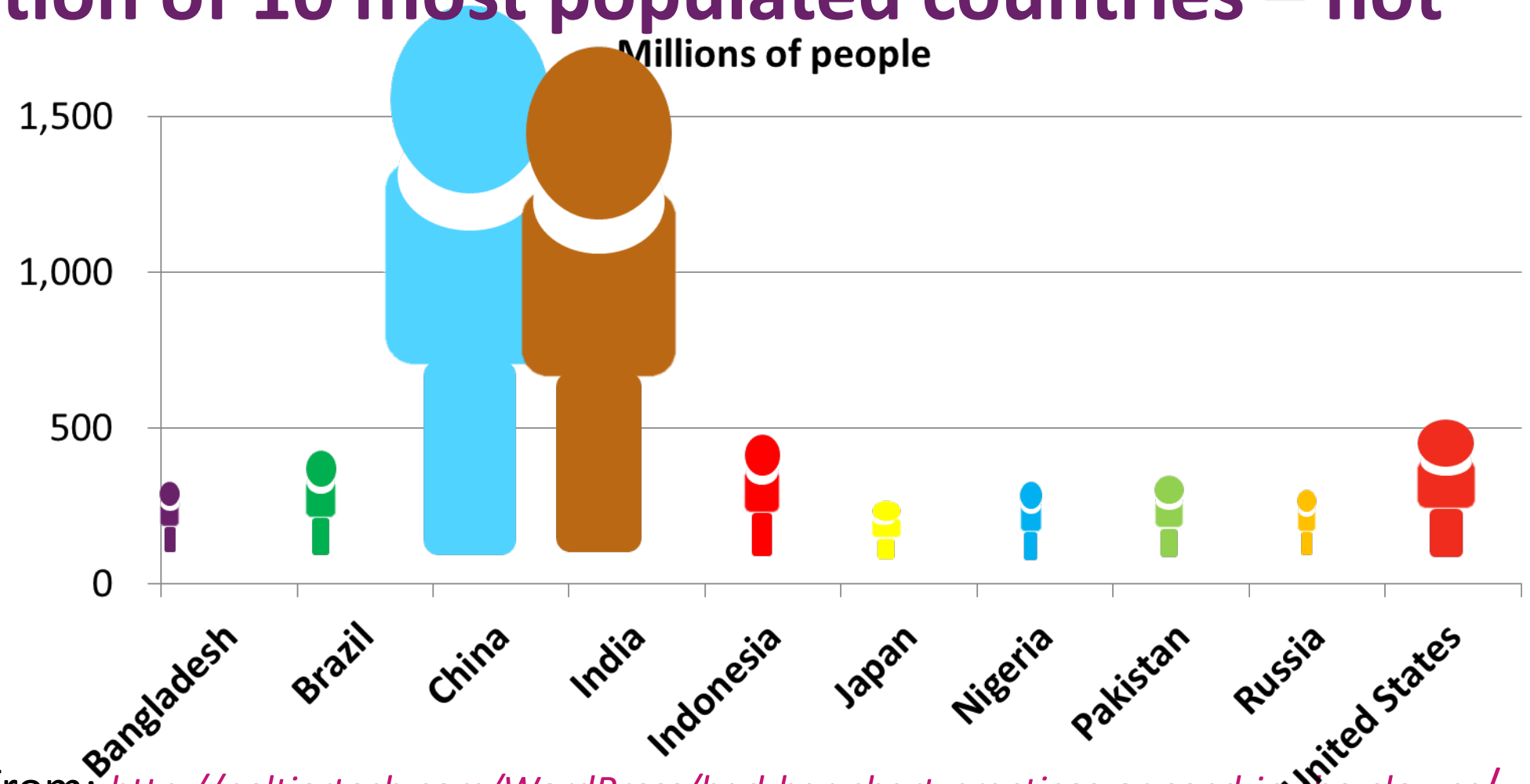


# … 3D barrels misrepresent data

- Smallest barrel in largest barrel – more than 12 in one layer (more than 1200%)
  - A lot more if whole volume considered!
- Do not use 3D marks (barrels) to display 1D data (price).

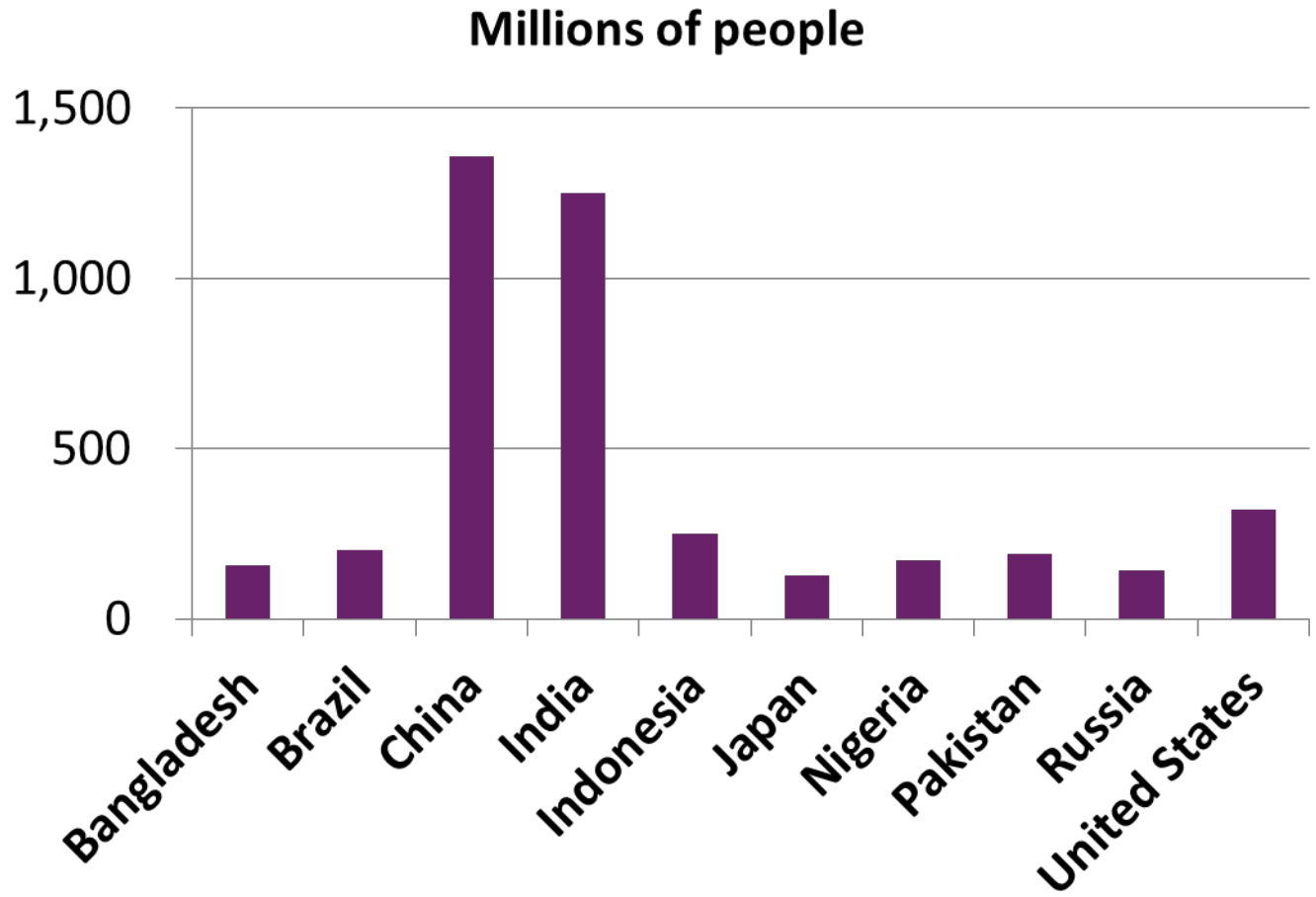


# Population of 10 most populated countries – not right!

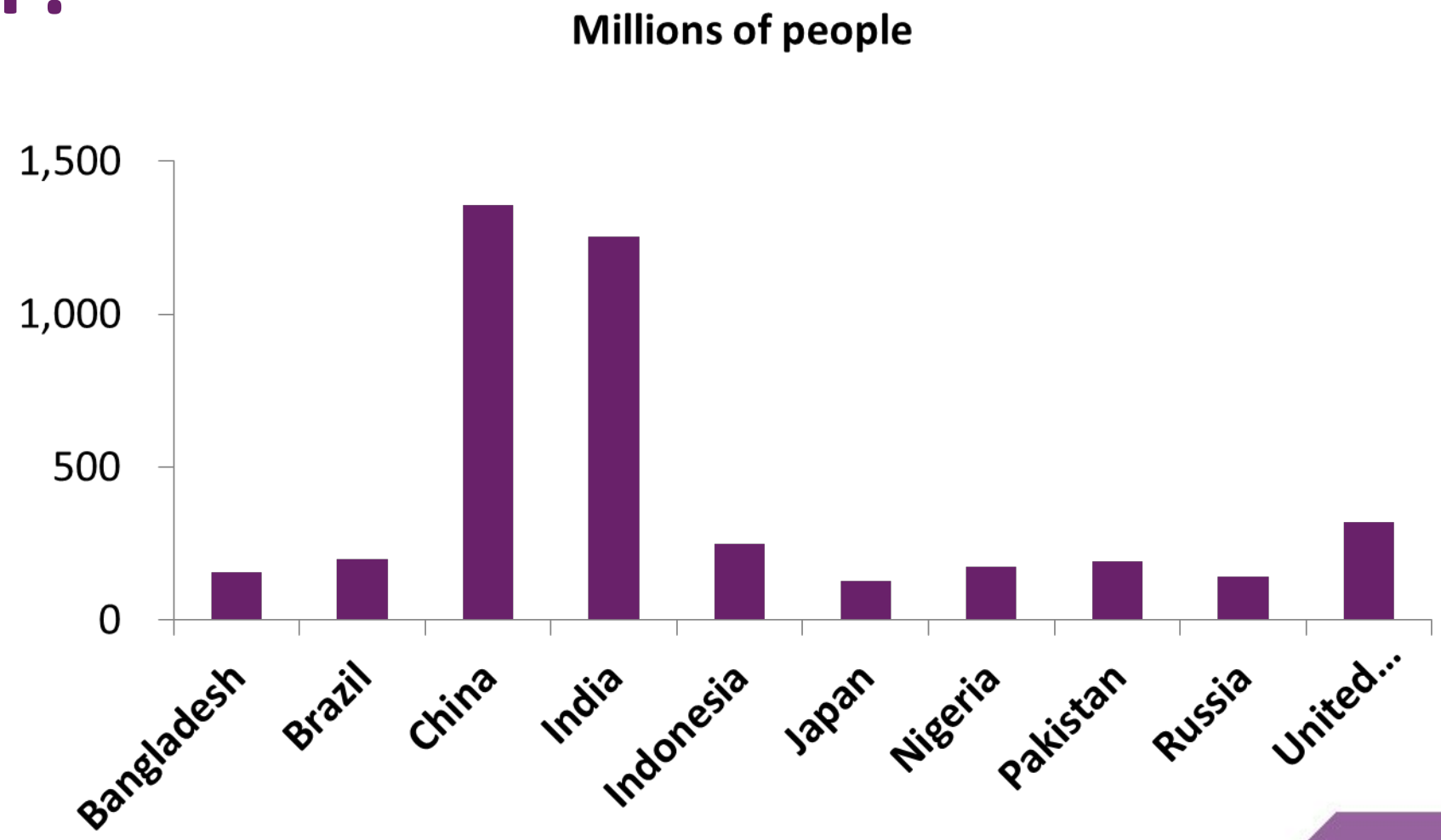



Adapted from: http://peltiertech.com/WordPress/bad-bar-chart-practices-or-send-in-the-clowns/ [accessed 29/01/2019]. Also next few slides.

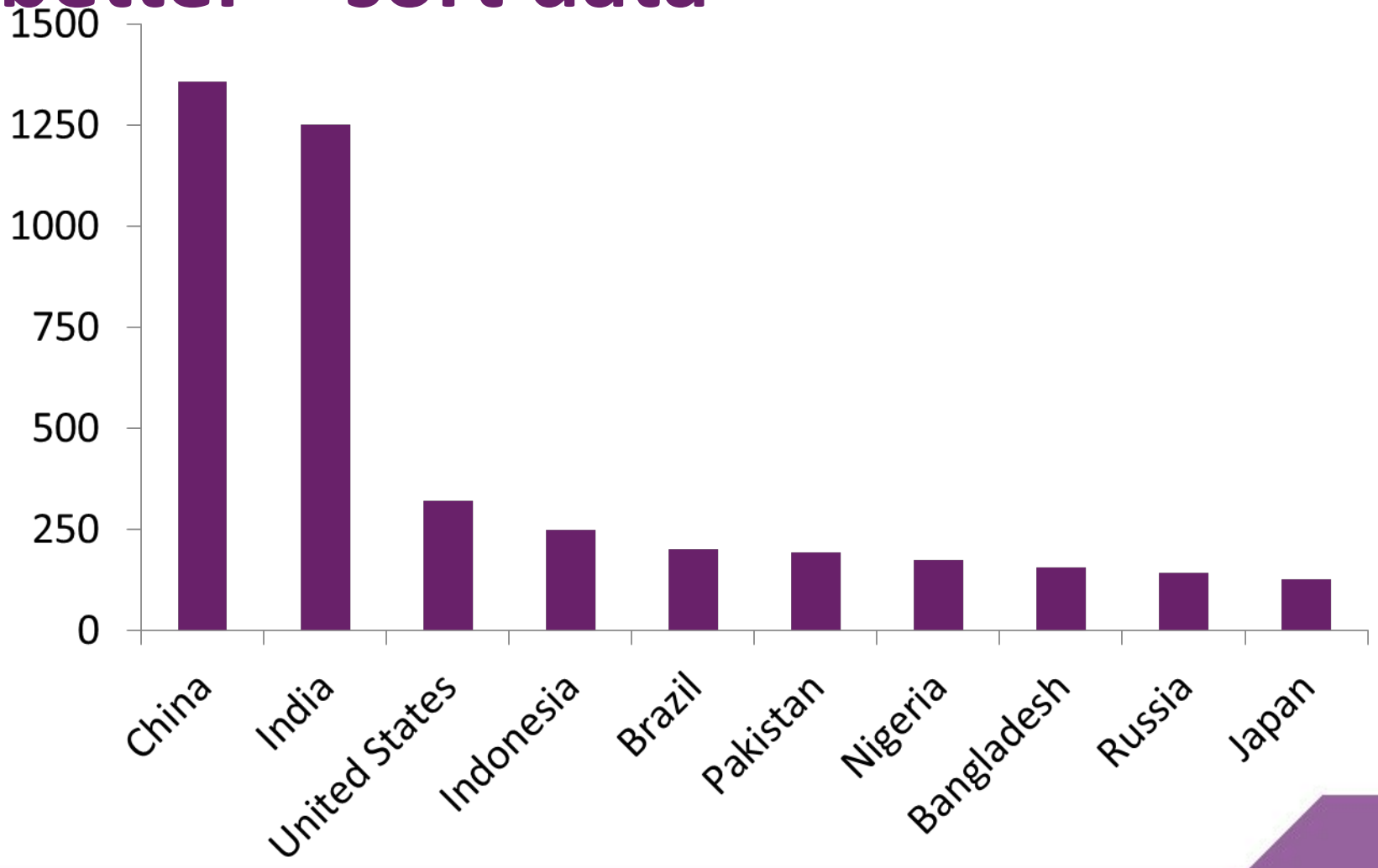
# Better


Millions of people
1,500
1,000
500
0
Bangladesh
Brazil
China
India
Indonesia
Japan
Nigeria
Pakistan
Russia
United States

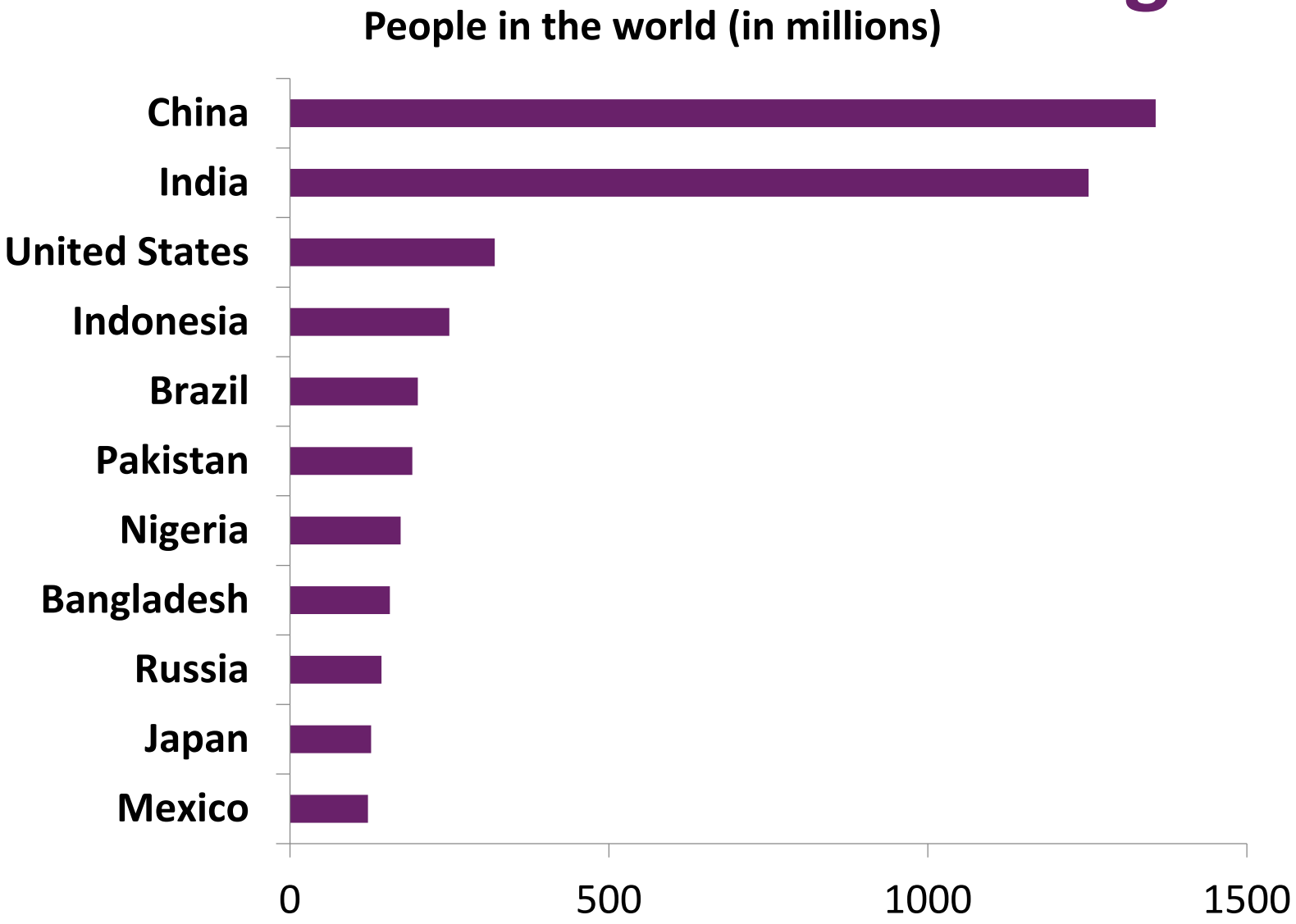
# Better?


Millions of people
1,500
1,000
500
0
Bangladesh
Brazil
China
India
Indonesia
Japan
Nigeria
Pakistan
Russia
United...

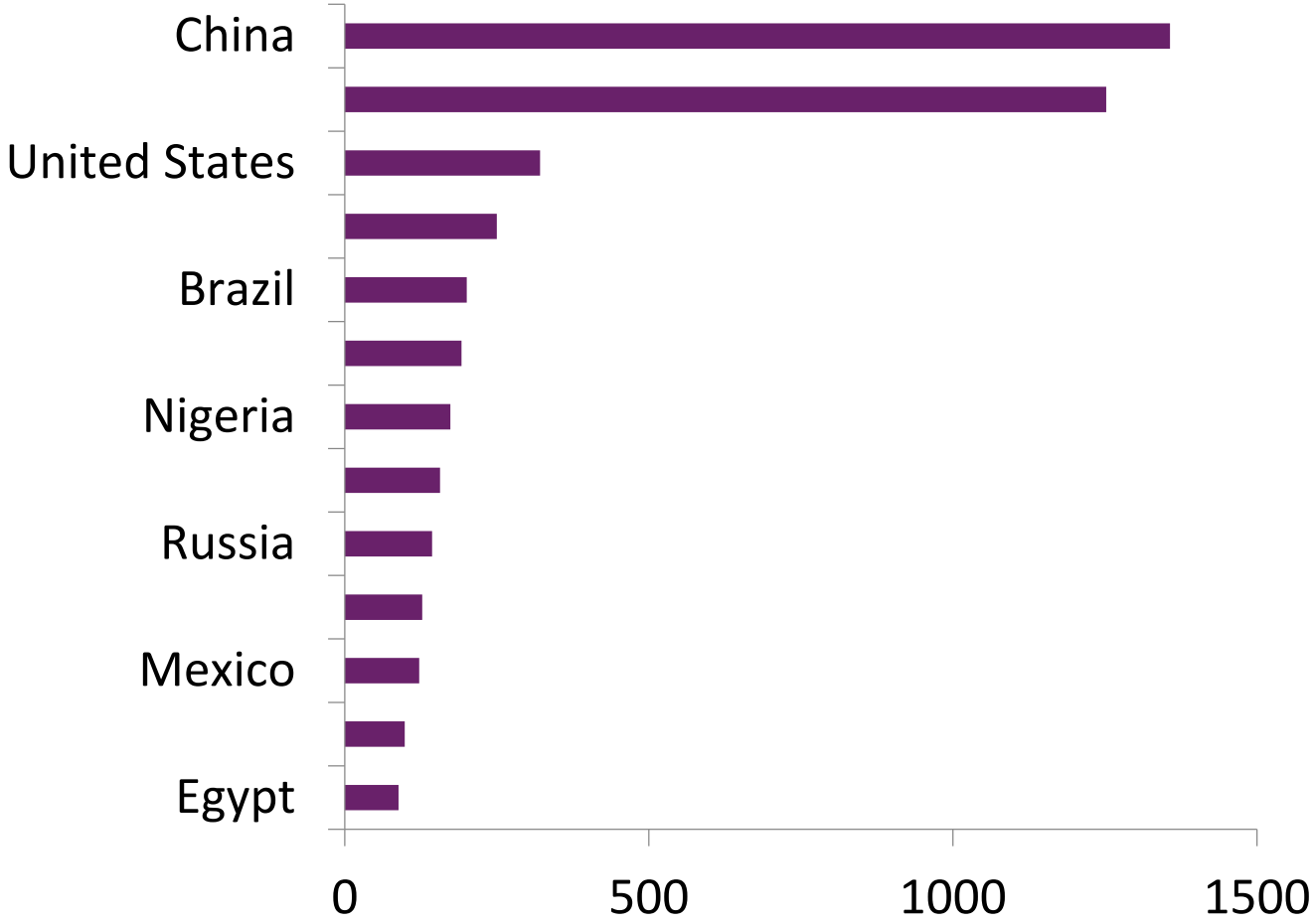
# Even better – sort data



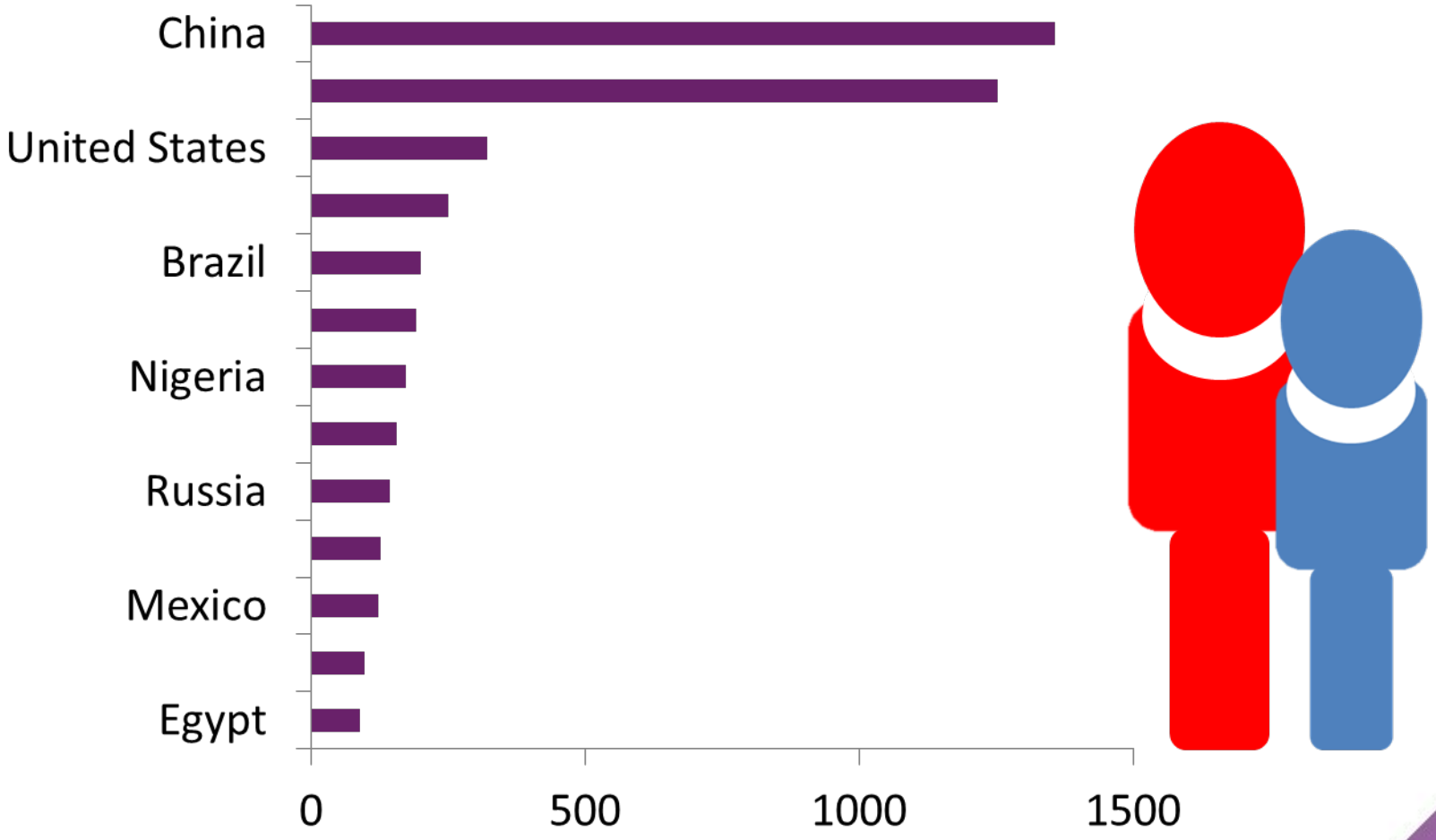
# Better if needing to scroll



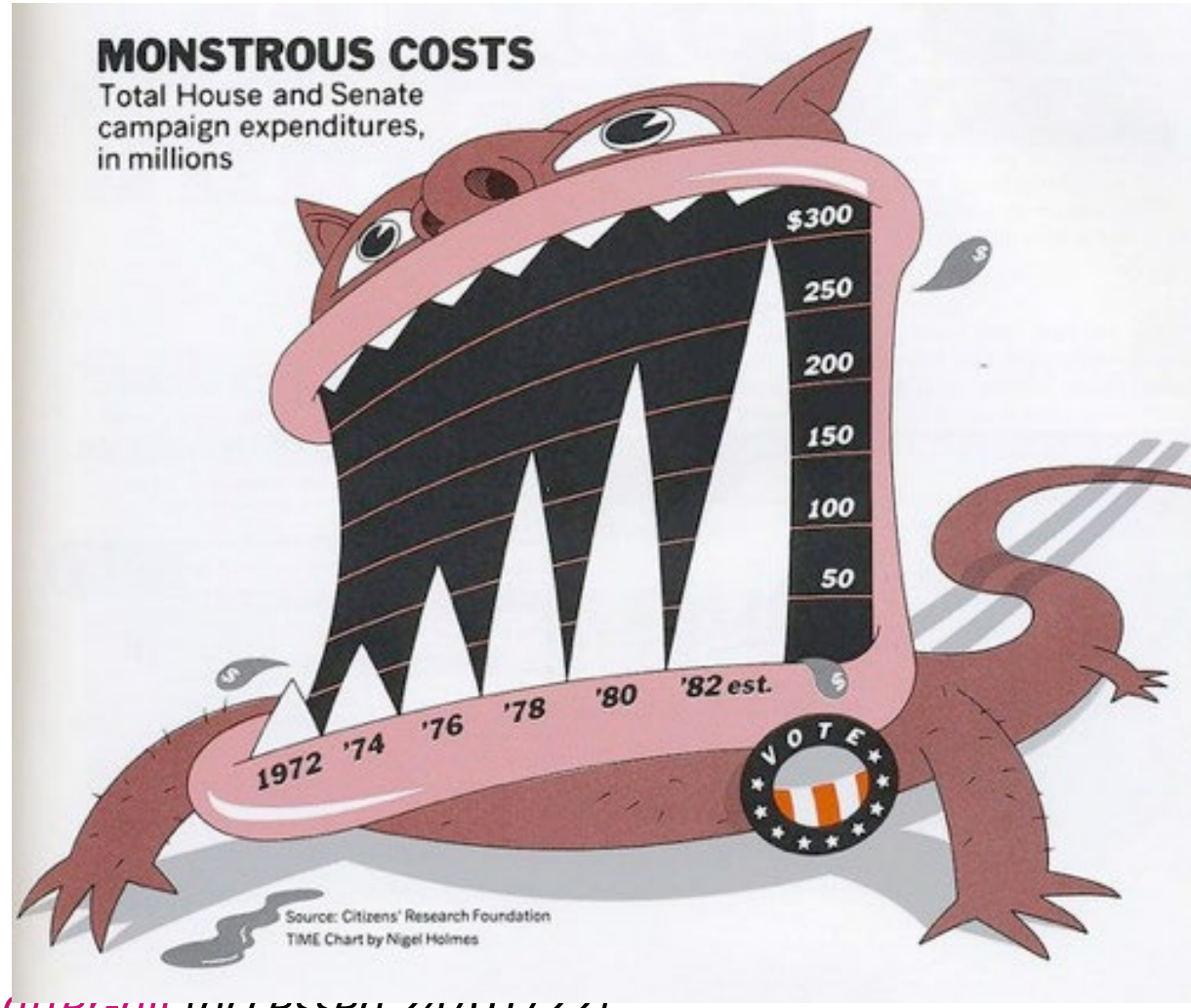
# … and it can be easily expanded (3 more countries)



# … but if you really want the people …



# Some designs distract from data?
# Nigel Holmes designs




https://eagereyes.org/criticism/chart-junk-considered-useful-after-all [accessed 24/01/22]

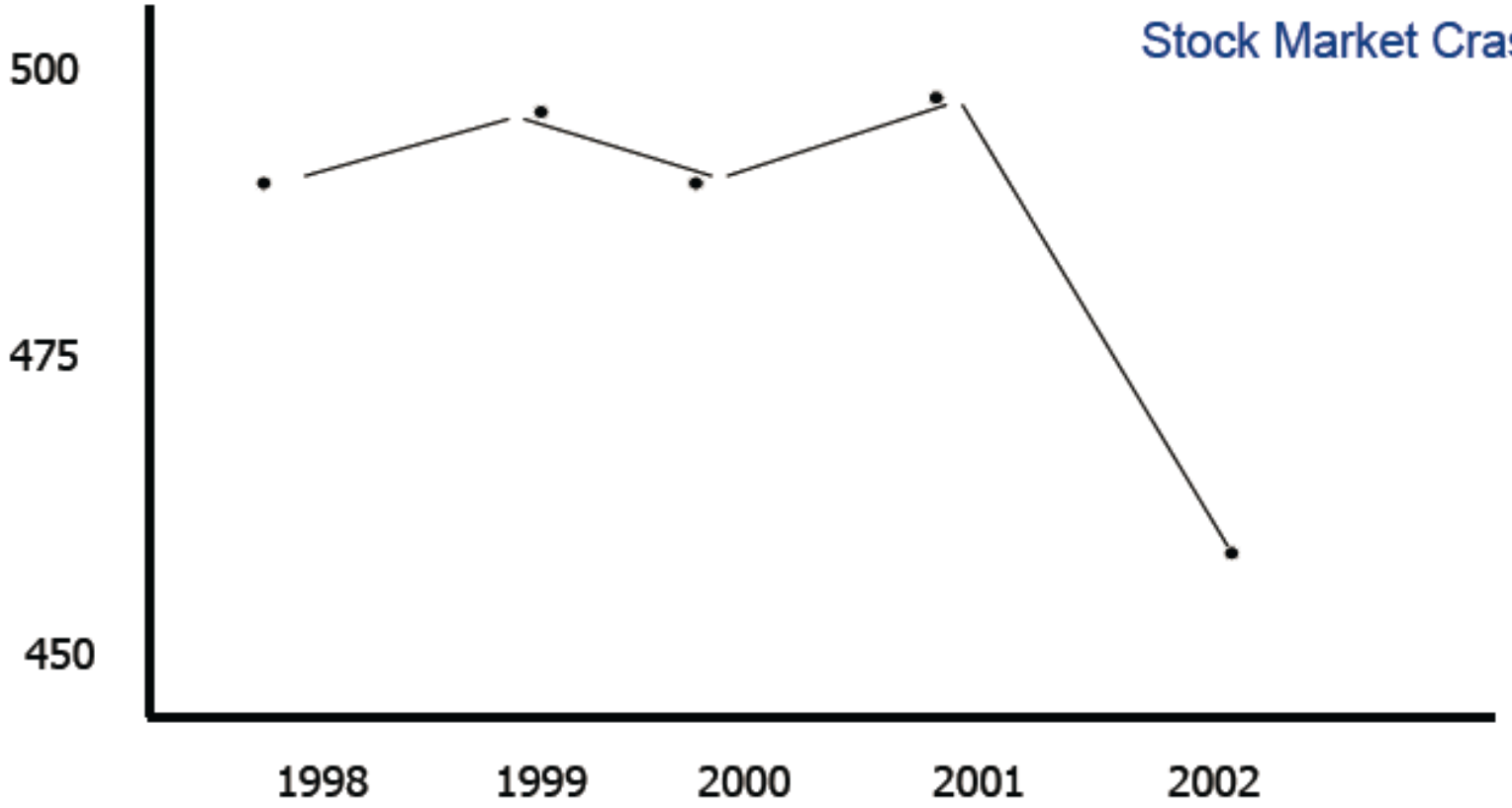
# Are visual ornaments a problem?

- Do they prevent us from understanding the data?
- Do they convey additional information?
- Experimental tests suggest
  - No significant difference between plain and ornated visualisation
  - No significant difference in recall accuracy after a 5 min break
  - Better recall for Holmes charts of the data topic and the details after a 3 week break
  - People were better at identifying messages in the Holmes visualisations.
  - People preferred Holmes charts
- Problems:
  - Space use
  - Interpretation of data
  - Trust – biased image?

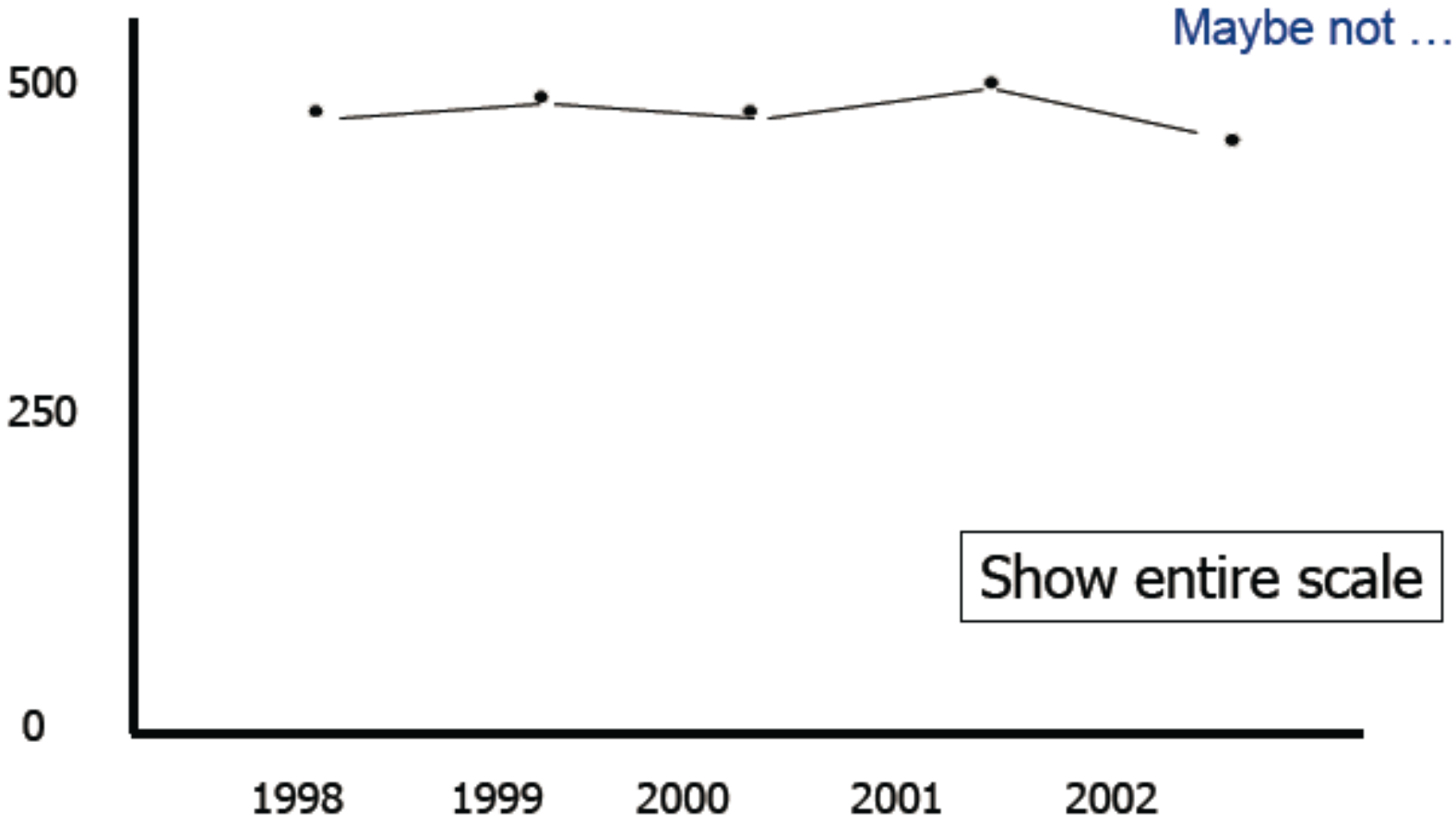
# Content (4)

- Data
- Why visualise?
- Requirements
- Problem visualisations
- **Tips for visualisation**
- Summary

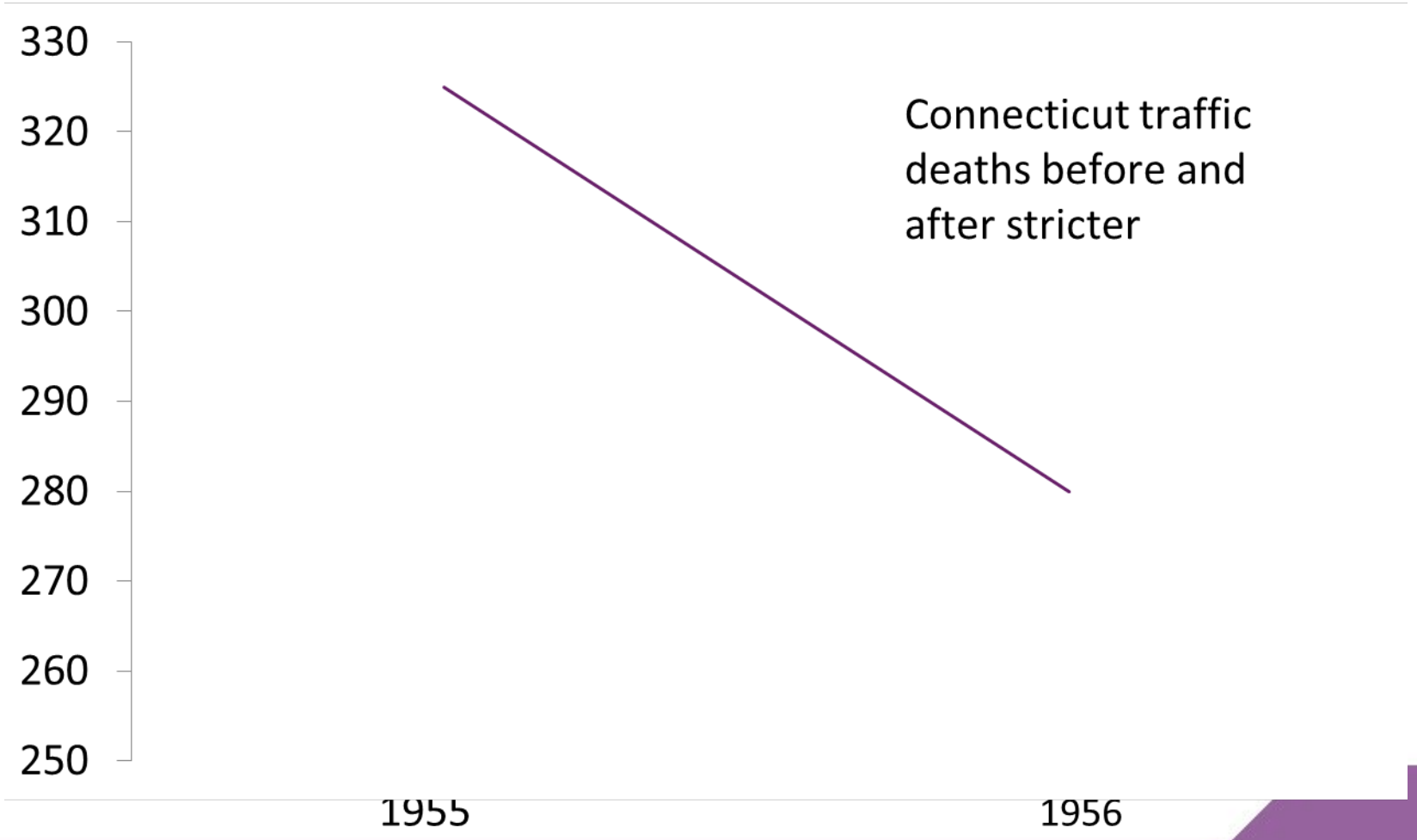
# Show context to ensure graphical integrity



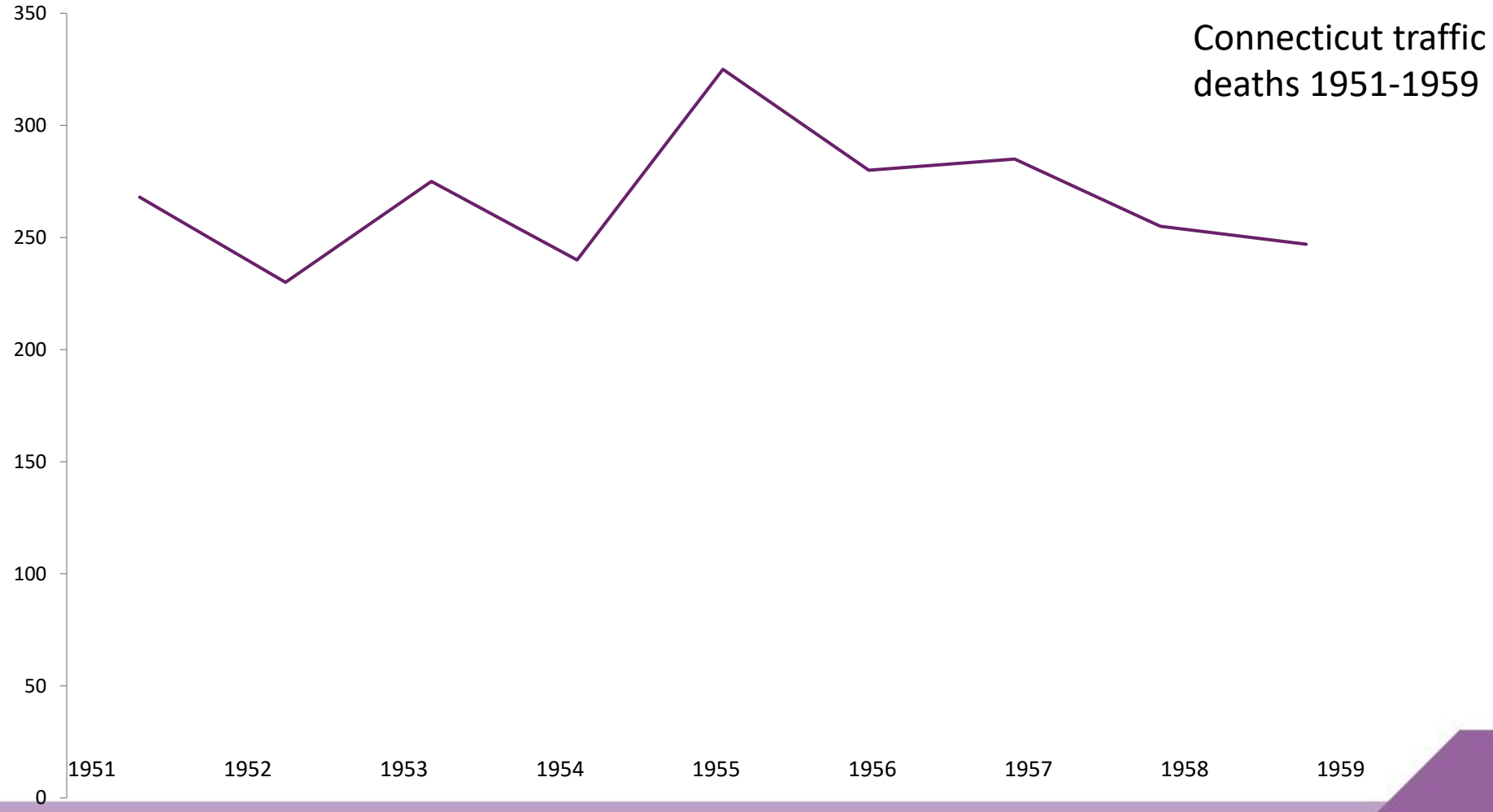
# … show context



# Other example – show context



# Better plot


Connecticut traffic deaths 1951-1959
350
300
250
200
150
100
50
0
1951
1952
1953
1954
1955
1956
1957
1958
1959

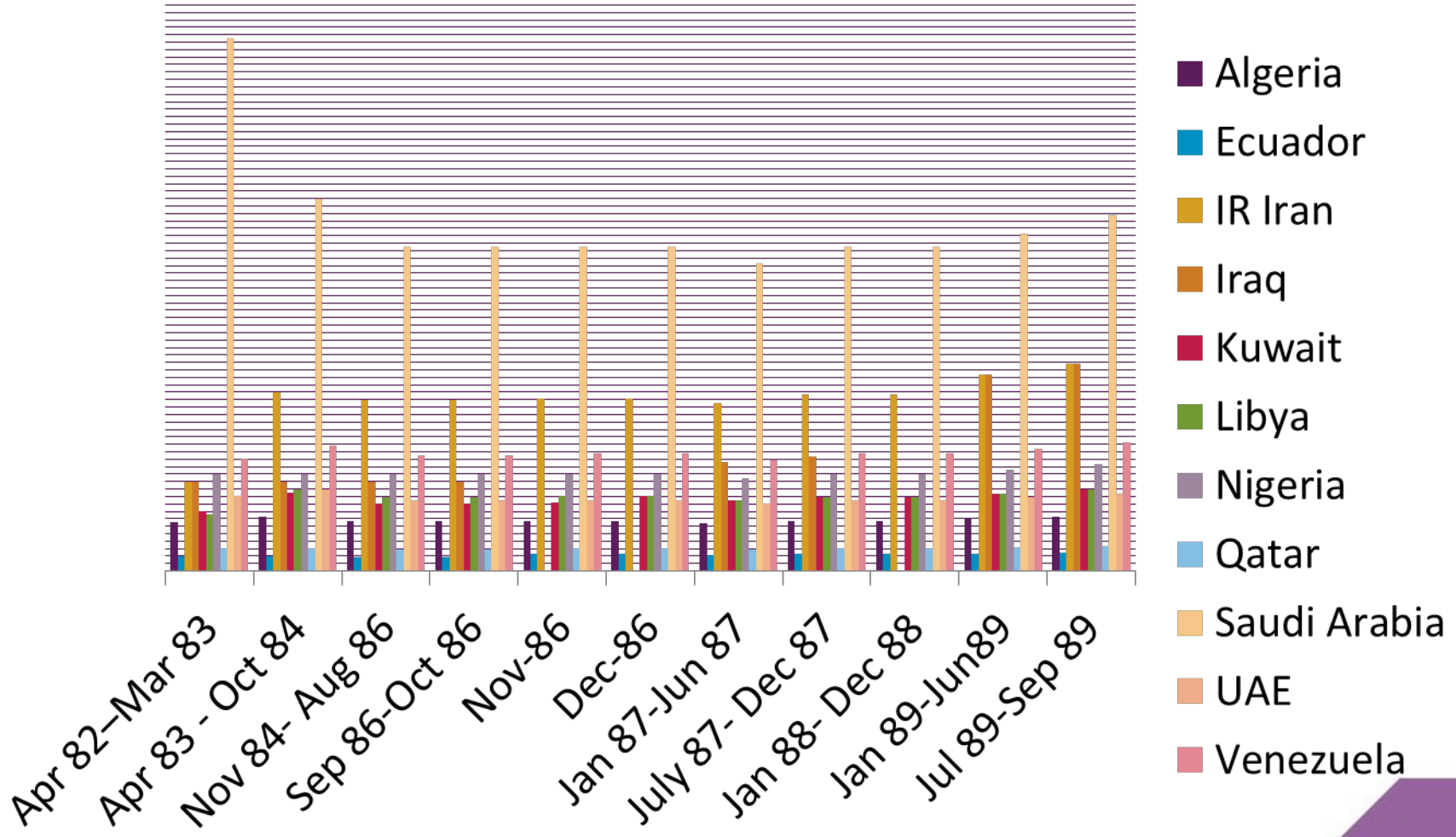
# Maximise data-ink ratio

- $dataInkRatio = \frac{dataInk}{totalInkUsedInGraphic}$

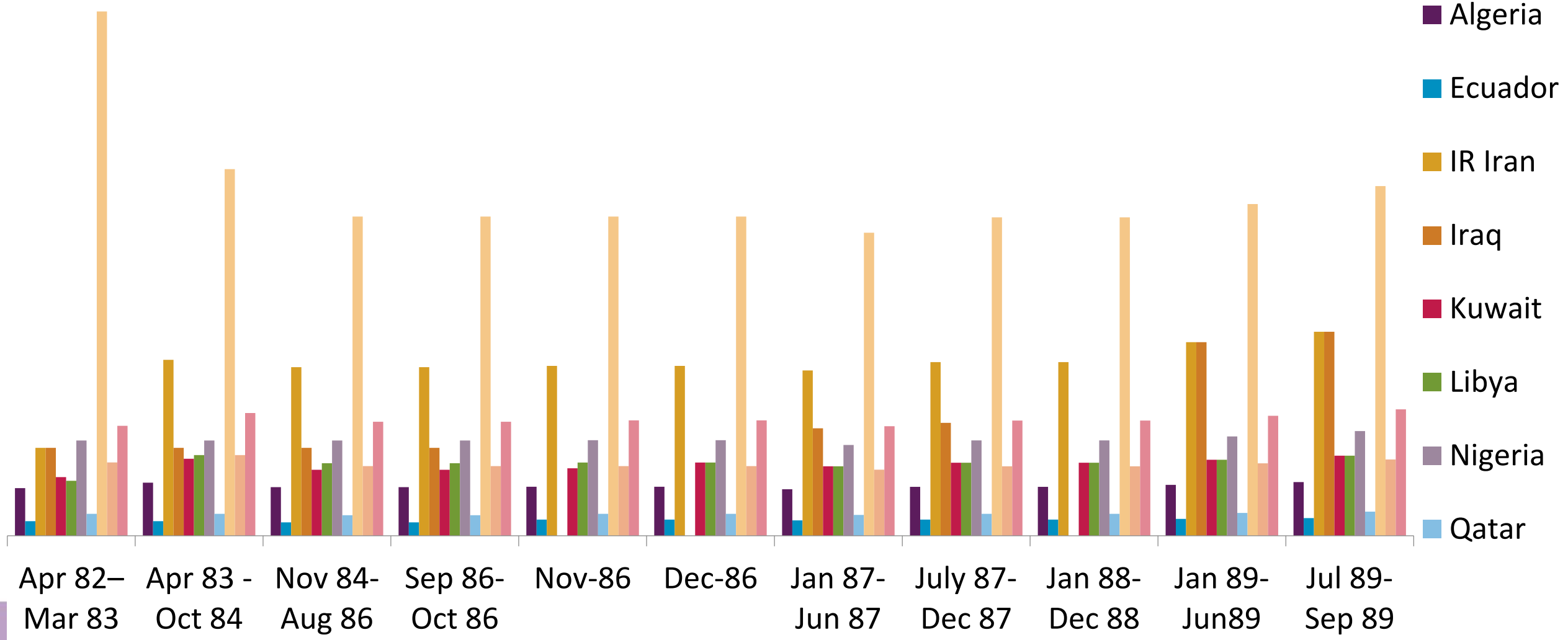
- Most of the ink should be used to present the data
  - NOT to ornate the data

- Retain only what cannot be deleted without loosing information.

# Wrong data-ink ratio



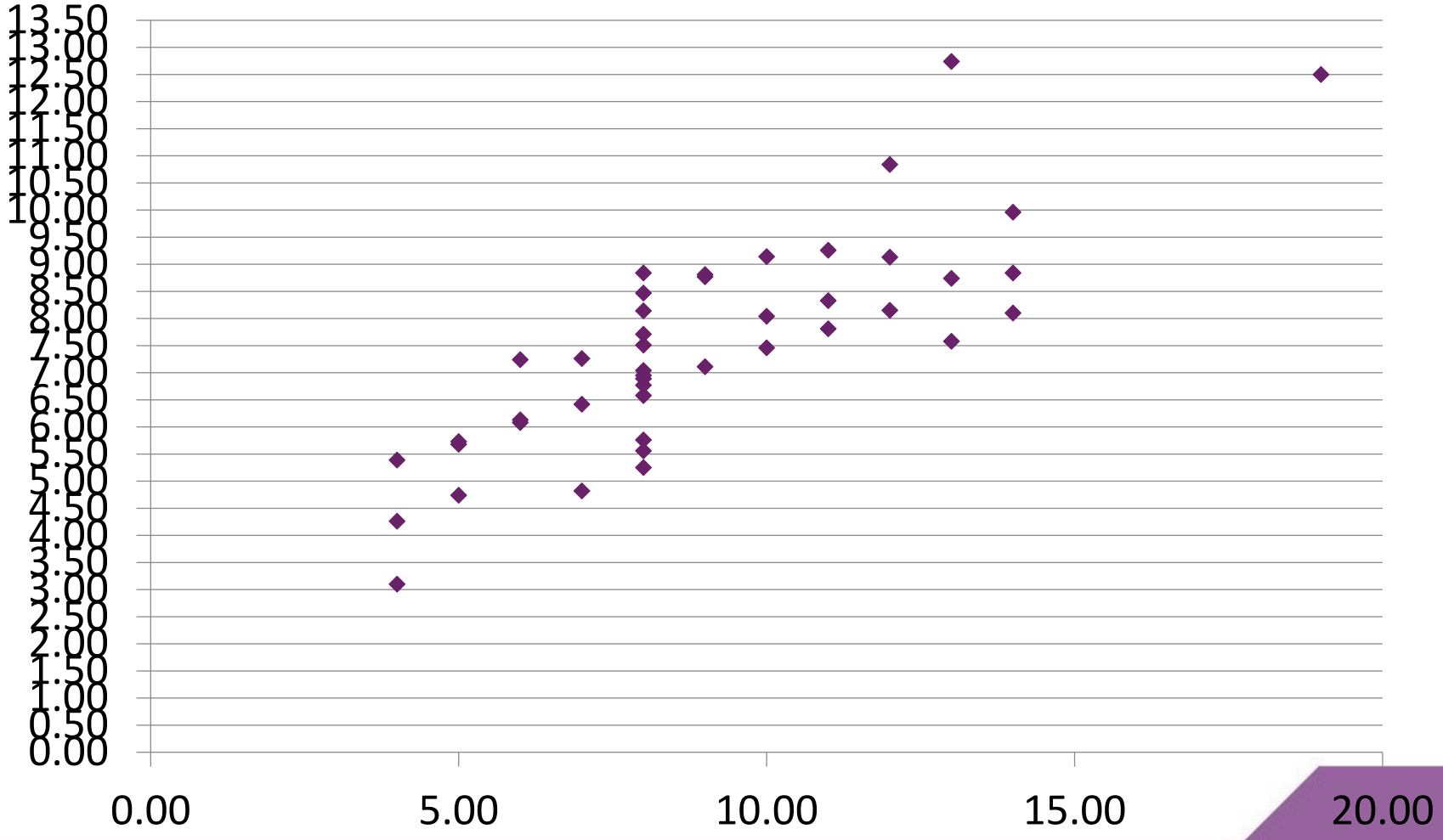
# Better bar plot



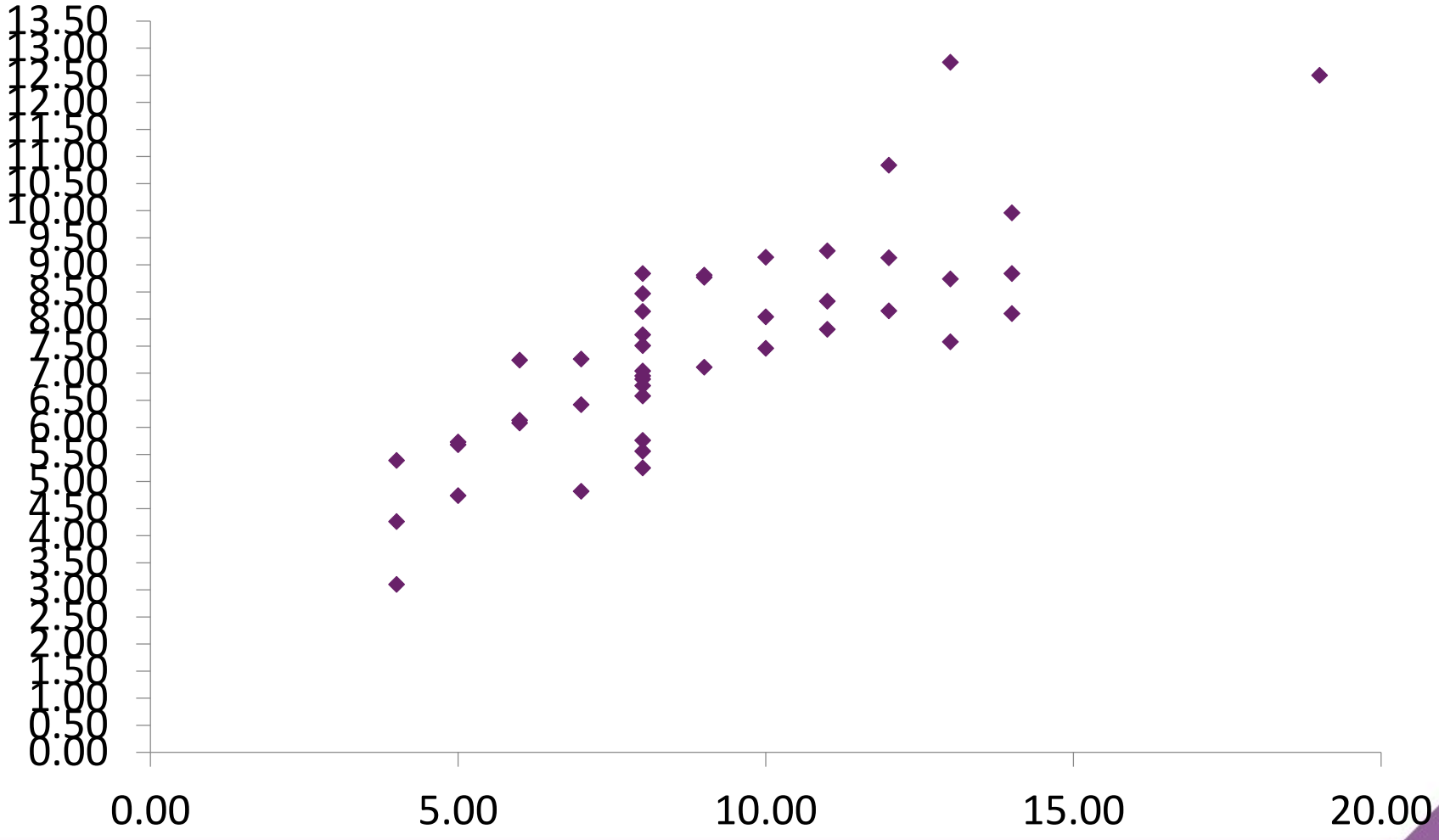
# Data-ink ratio

- Delete non-data ink
  - Within reason!!!!
- What is each bit of ink telling?
  - Each bit of ink requires a reason.
- Delete redundant information
- Delete decorations

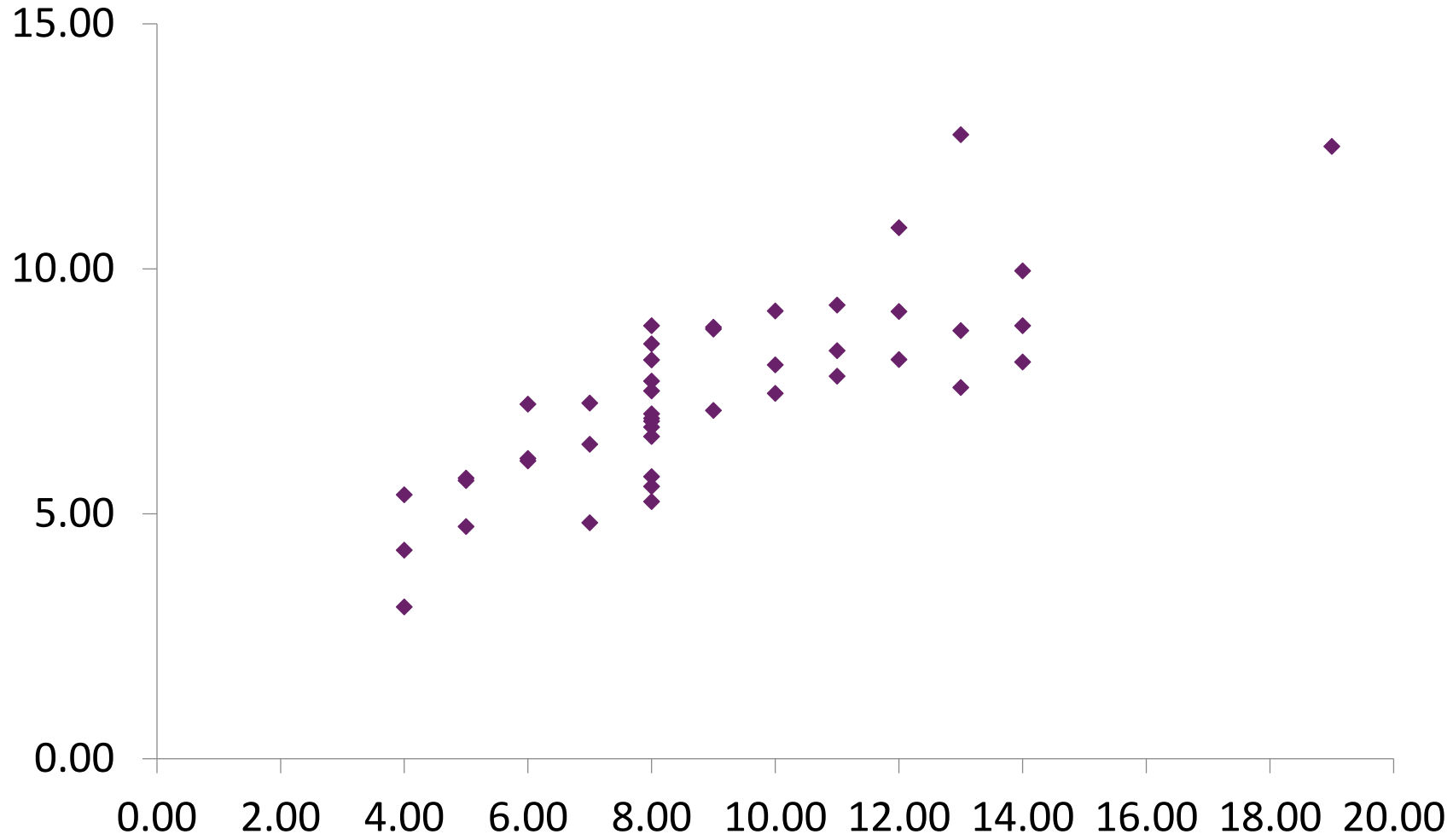
# Poor data-ink ratio



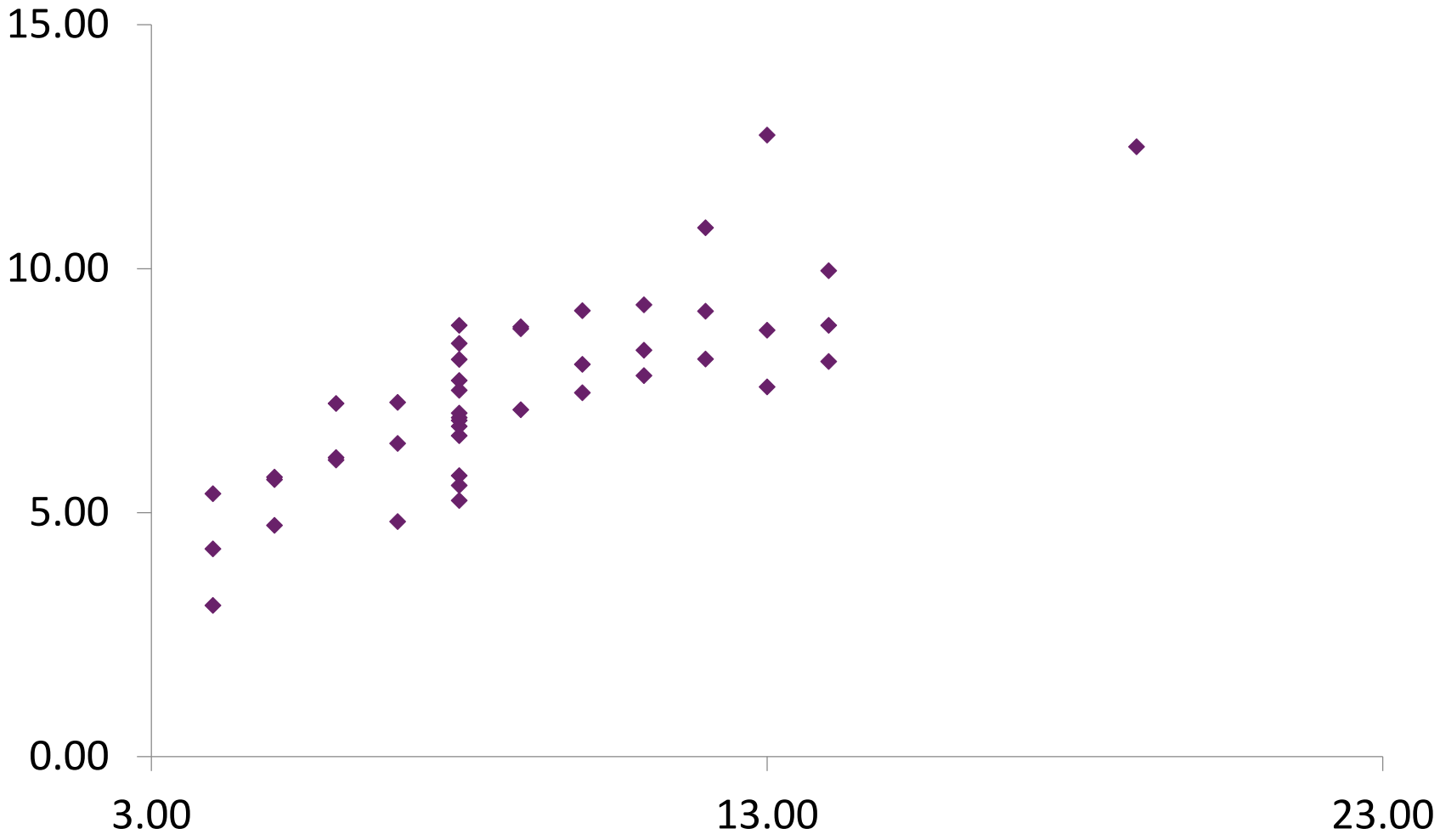
# Remove horizontal lines



# Adjust Y axis



# Adjust X axis (note it does not start at 0).




Removal of points is not always useful.
It may be easier to understand if it starts at zero.

# Other

- Solid diamond markers could be changed for unfilled circles
  - This would allow to see any overlapping data

# Very large datasets

- If there is too much data
  - Consider data summarisation prior to visualisation
  - Do you need to present all the data?
  - Do you need to consider presenting the data at several different levels of detail?
    - E.g. OPEC production allocations by periods vs. summary of production allocations.

# Good visualisations chracteristics

- Simplicity – make graphs and tables as simple as possible
- Graphs are gen. better than tables
  - Except when the amount of data is small
- Titles of visualisations should be meaningful
- *Explain* the graph
  - X variable – including unit of measure
  - Y variable – including unit of measure
  - Scale and limitations
  - Include a meaningful title

# … good visualisations

- Number representation should be directly proportional to the actual amounts being represented
  - lie factor should be 1
- Use labels to explain any potential graphical mislead or ambiguity
- Show changes in data NOT in design
- Present cause and effect in the visualisation

# Summary

- Show data
  - Emphasis should be on the information contained in the data NOT on graphics.
- Avoid presenting data in a misleading way
  - Consider
    - Lie factor
    - Scale
- Make large data sets coherent
  - Data summarisation may be needed
- Maximise data-ink ratio
  - Within reason
- Visualisation should encourage the viewer compare different pieces of data