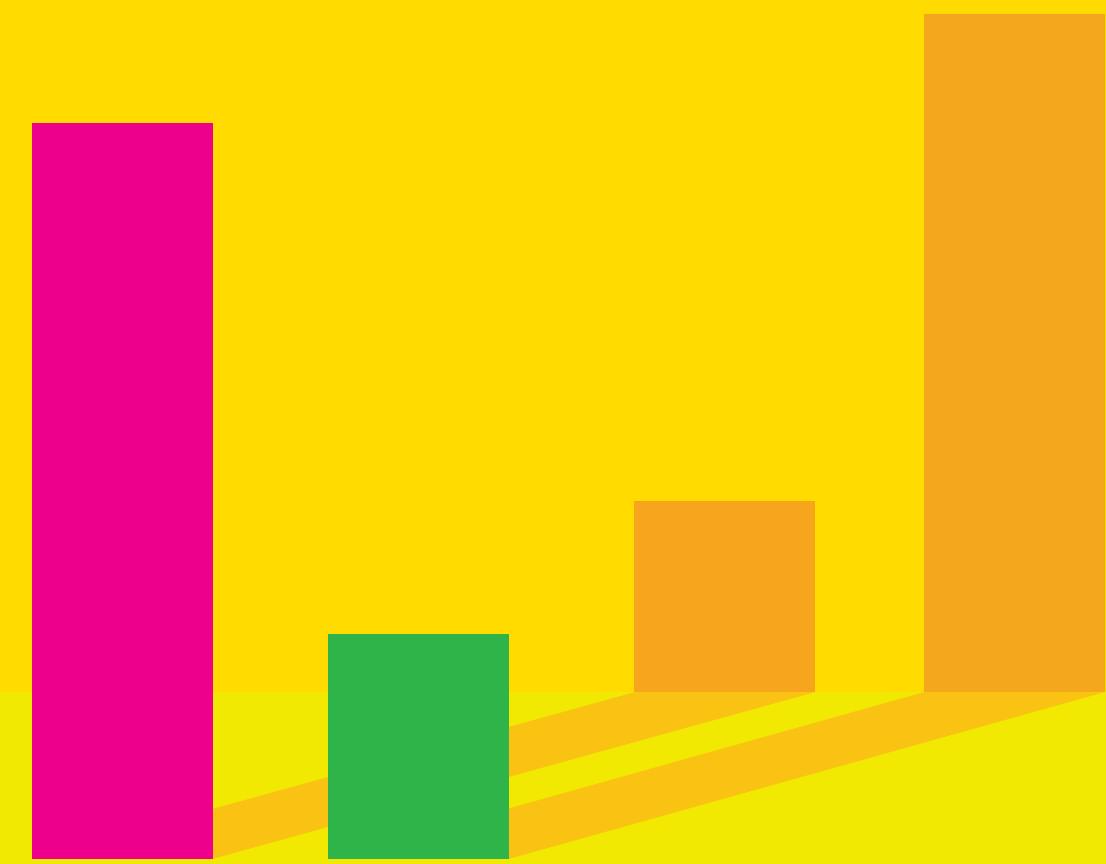


HOW TO MAKE DECISIONS IN DATA VISUALIZATION



Alberto Cairo

Banco de Portugal, 2020

Time spent in a life of a Data Scientist

@datavizzdom

Gulrez

Perception



Creating ML
Models

Time spent in a life of a Data Scientist

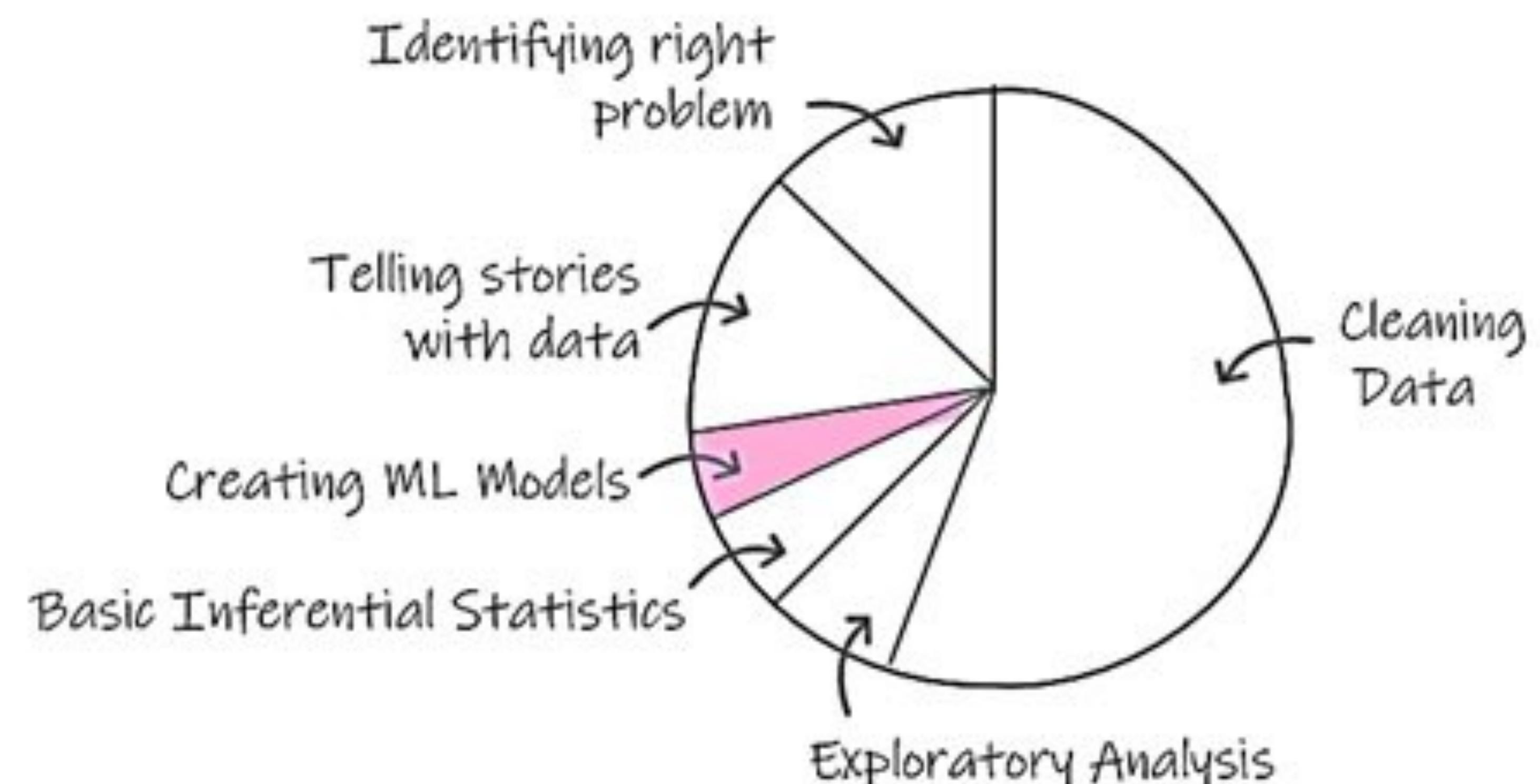
@datavizzdom

Gulrez

Perception



Reality



Time spent in a life of a Data Scientist

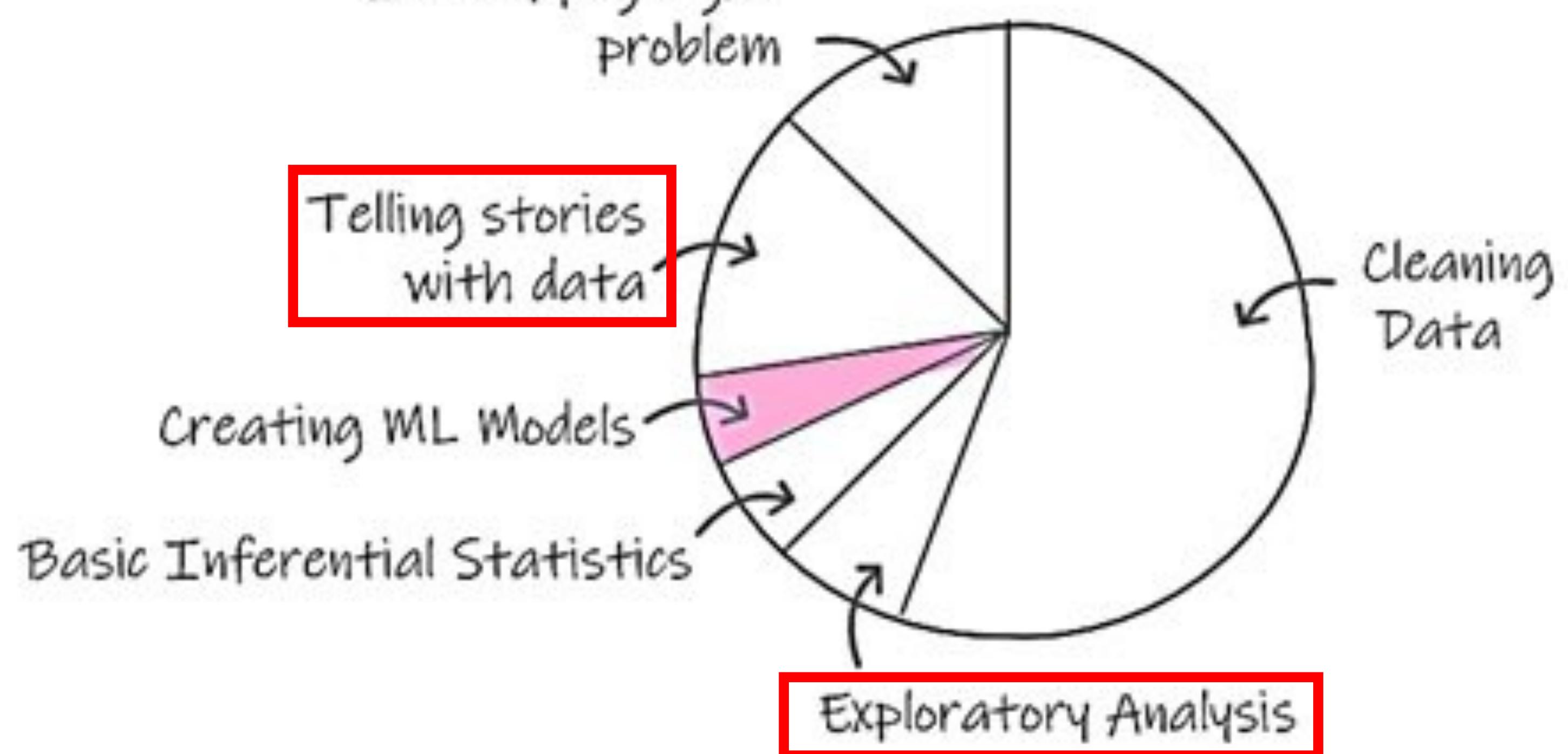
@datavizzdom

Gulrez

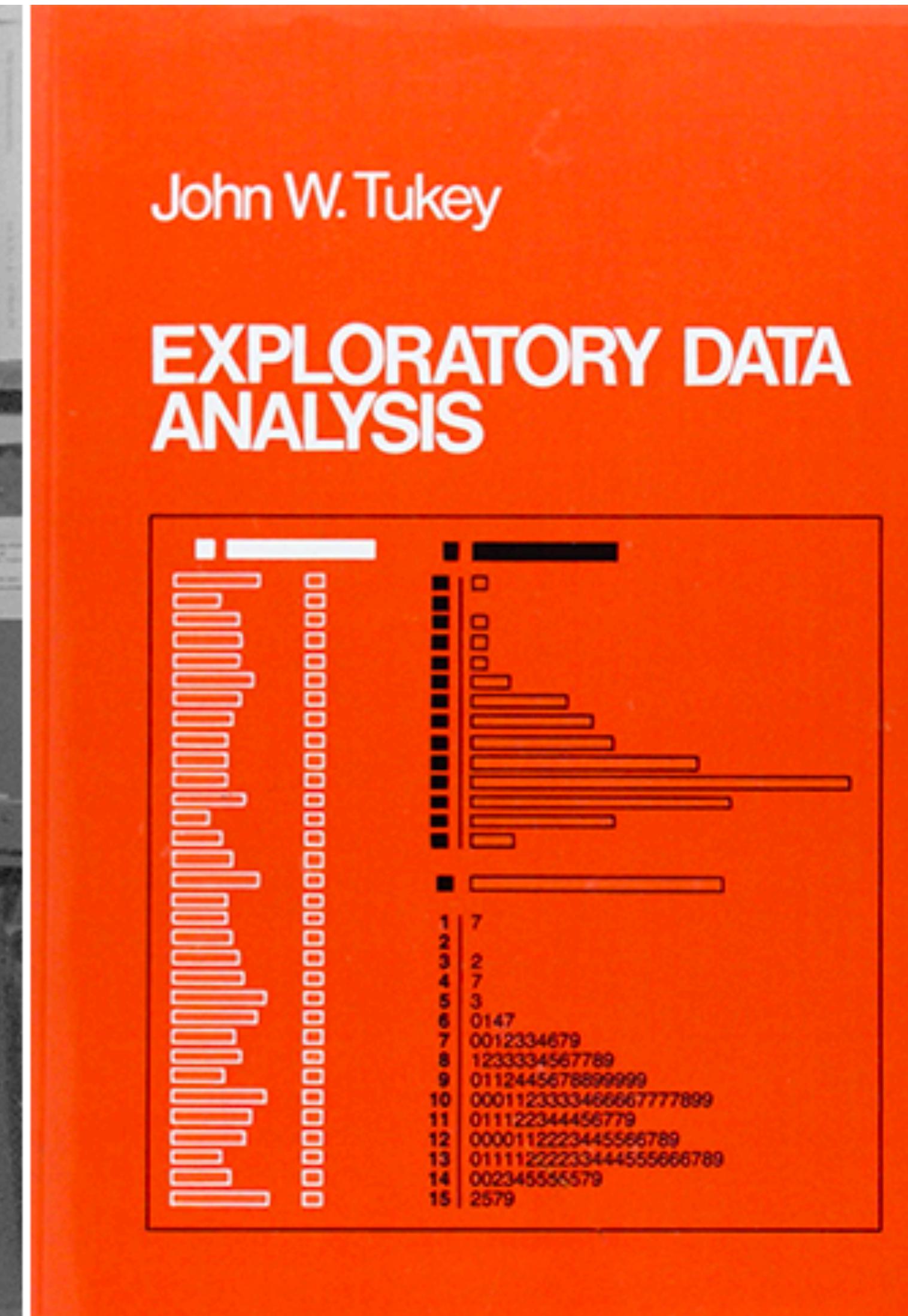
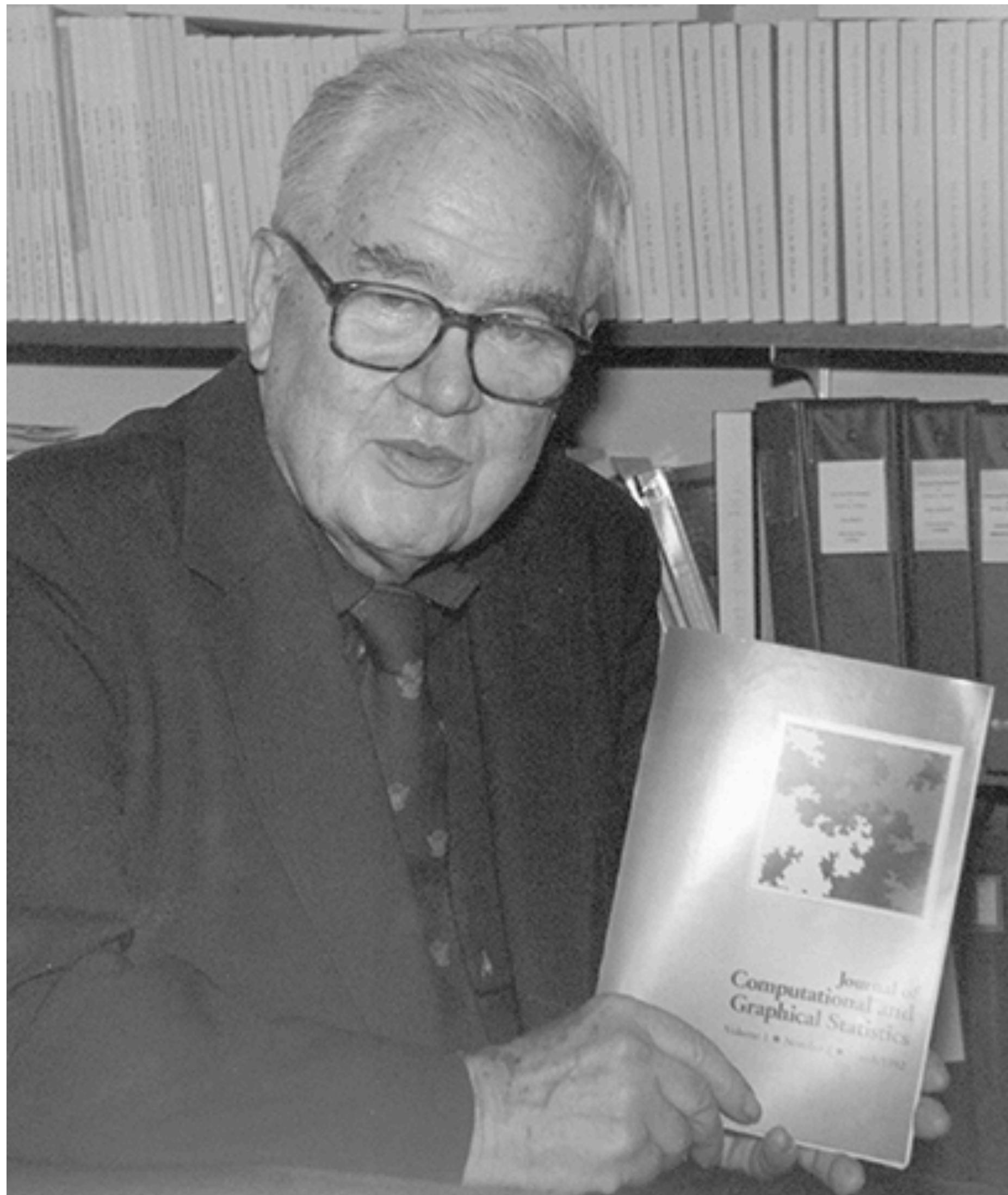
Perception



Reality



We are living in a Golden Age of visualization



“The greatest value
of a picture is when it
forces us to notice
what we never
expected to see.”

John W. Tukey

We are living in a Golden Age of visualization

More info ➔

More info ➔



Sales per countries



Top 10 products

430



We are living in a Golden Age of visualization

Ed Hawkins's 'Warming stripes' (read more: <https://chezvoila.com/blog/warmingstripes/>)



We are living in a Golden Age of visualization

The Washington Post
Democracy Dies in Darkness

Sections 

Alberto Cairo To... 

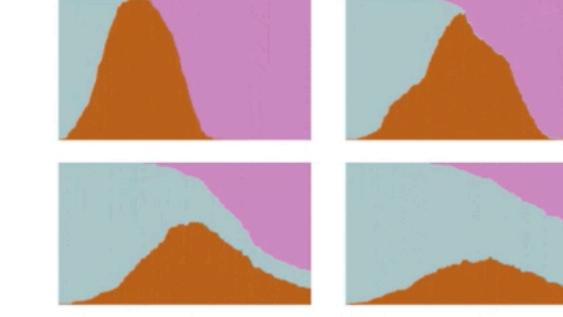








2.6k



Health

Why outbreaks like coronavirus spread exponentially, and how to “flatten the curve”

By Harry Stevens March 14, 2020

PLEASE NOTE

The Washington Post is providing this story for free so that all readers have access to this important information about the coronavirus. For more free stories, [sign up for our daily Coronavirus Updates newsletter](#).

<https://www.washingtonpost.com/graphics/2020/world/corona-simulator/>

We are living in a Golden Age of visualization

The Post's visual journalism, which involves staff throughout the newsroom, has attracted large audiences and contributed to record subscriber growth.

Six of the seven most visited stories in The Washington Post's history have been graphics, including the [coronavirus simulator](#) that became the most visited article in The Post's history, with more than three times as many visits as the second. It also includes this year's [Democratic candidate quiz](#), which set the record for converting readers to subscribers.

<https://www.washingtonpost.com/pr/2020/06/26/washington-post-expand-graphics-design-teams-with-14-new-positions/>

We are living in a Golden Age of visualization

FLORIDA: WHY ARE SO MANY SENIORS STRUGGLING?

Asset Limited, Income Constrained, Employed (ALICE) is a segment of the U.S. population who do not meet federal poverty levels but are struggling to make ends meet. In Florida, households 65 years and older saw the greatest increases below the ALICE threshold across all ethnic and racial groups; however, total households for those 25 years and under decreased. Florida also leads the nation with the greatest number of seniors

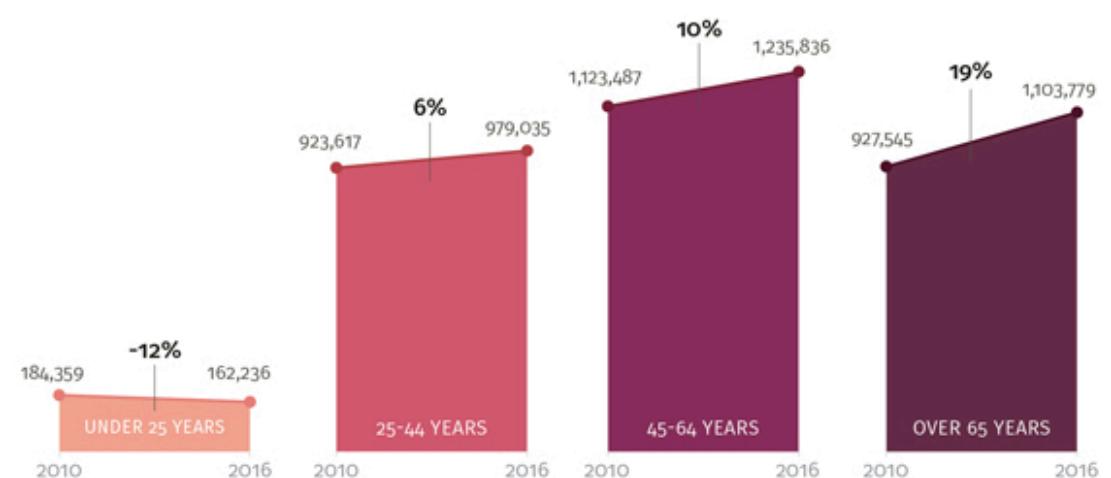
with many dependent on Social Security as their primary source of income. Housing is at an all-time high negatively affecting the overall health of many low-income Floridians. With formal care costs out of reach for most households, informal caregivers will continue to feel the financial burden of long term care. As more Floridians struggle to get by, will Florida have policies in place to address the monumental impacts of an aging population?

NEARLY 50% OF FLORIDIANS FIGHT TO SURVIVE



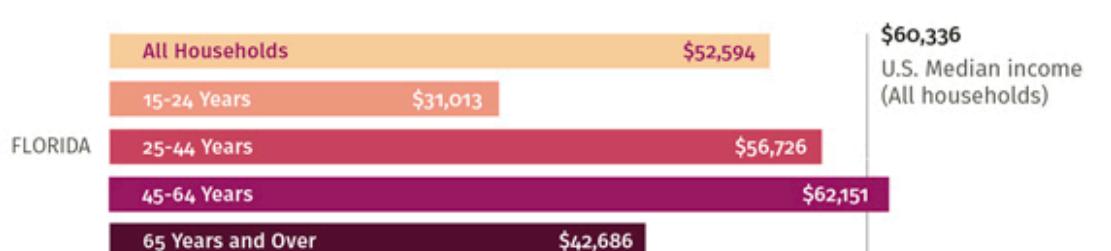
SENIORS, 45-64 YEAR OLDS INCREASINGLY STRUGGLE

PERCENT CHANGE OF HOUSEHOLDS BY AGE BELOW THE ALICE THRESHOLD FROM 2010 TO 2016



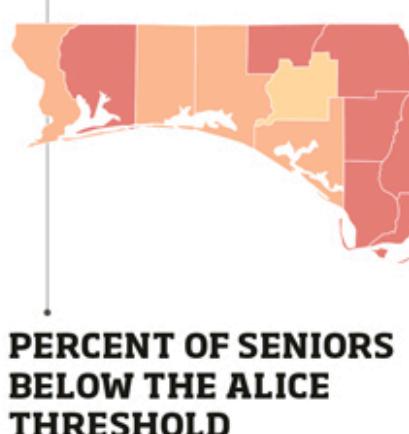
MEDIAN INCOME BARELY COVERS THE BASICS

THE STRUGGLE TO AFFORD THE ESSENTIALS LEAVES LITTLE ROOM FOR UNEXPECTED COSTS



FLORIDA: THE NATION'S OLDEST STATE

MAINE, WEST VIRGINIA AND VERMONT FOLLOW WITH THE HIGHEST PERCENTAGE OF SENIORS (65+)



HIGHEST PERCENTAGE OF ALICE HOUSEHOLDS 65+

Lafayette 49%

Glades 46%

LOWEST PERCENTAGE OF ALICE HOUSEHOLDS 65+

Wakulla 19%

Collier 20%

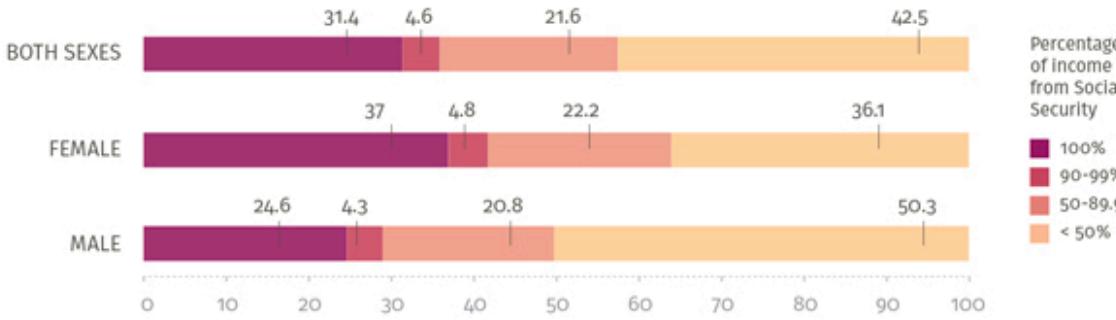
Leon 20%

20% OR LESS

50% OR MORE

SOCIAL SECURITY: A KEY INCOME SOURCE FOR SENIORS

RELIANCE ON SOCIAL SECURITY AS PERCENT OF TOTAL INCOME FOR PEOPLE 65+ YEARS



LONG TERM CARE STRAIN ON THE HORIZON

Costs for longterm care ("custodial care") support in Florida will continue to increase at alarming rates forcing vulnerable populations to seek alternatives. Medicare can cover a fraction of costs and the bulk of the financial burden falls on individuals and families. Medicaid is only available for Americans with the lowest incomes with caveats. When it comes to long term care, ALICE households are forced further to the margins.

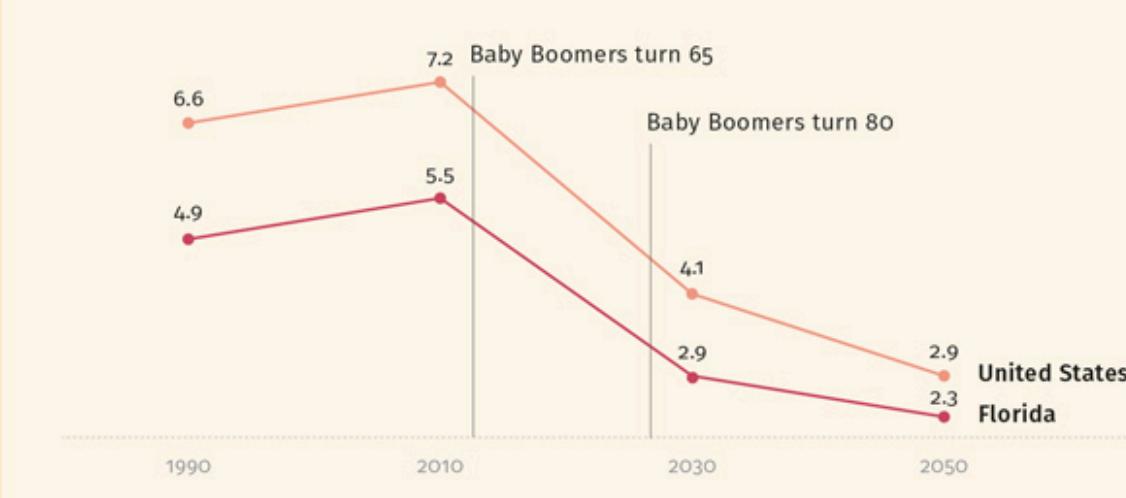
FORMAL CARE COSTS WILL CONTINUE TO OUTPACE U.S. INFLATION RATES

FLORIDA'S ANNUAL PROJECTED MEDIAN CARE COSTS IN THOUSANDS



NUMBER OF FAMILY CAREGIVERS PROJECTED TO PLUNGE

POTENTIAL CAREGIVERS AGED 45-64 YEARS FOR EACH PERSON AGED 80 AND OLDER



WHAT COULD FLORIDA DO TO HELP SENIORS?

ACCORDING TO AARP, COMPREHENSIVE PEOPLE-FOCUSED POLICIES COULD MAKE A BIG DIFFERENCE

- IMPROVE OVERALL LONG TERM SERVICES AND SUPPORT (LTSS):** Quality of life would increase, including the promise for seniors and adults with disabilities to afford housing.
- EXPAND HOME AND COMMUNITY-BASED SERVICES (HCBS):** More Floridians could avoid costly nursing homes and family caregivers would be able to receive assistance.
- EXPAND MEDICAID:** Non-elderly adults without dependents could be covered and the health coverage gap for nearly 400,000 Floridians would be reduced.

ANYONE can learn to make data visualizations.

Example of my students' work:

<https://www.deb.is/>

How to teach or learn data visualization?

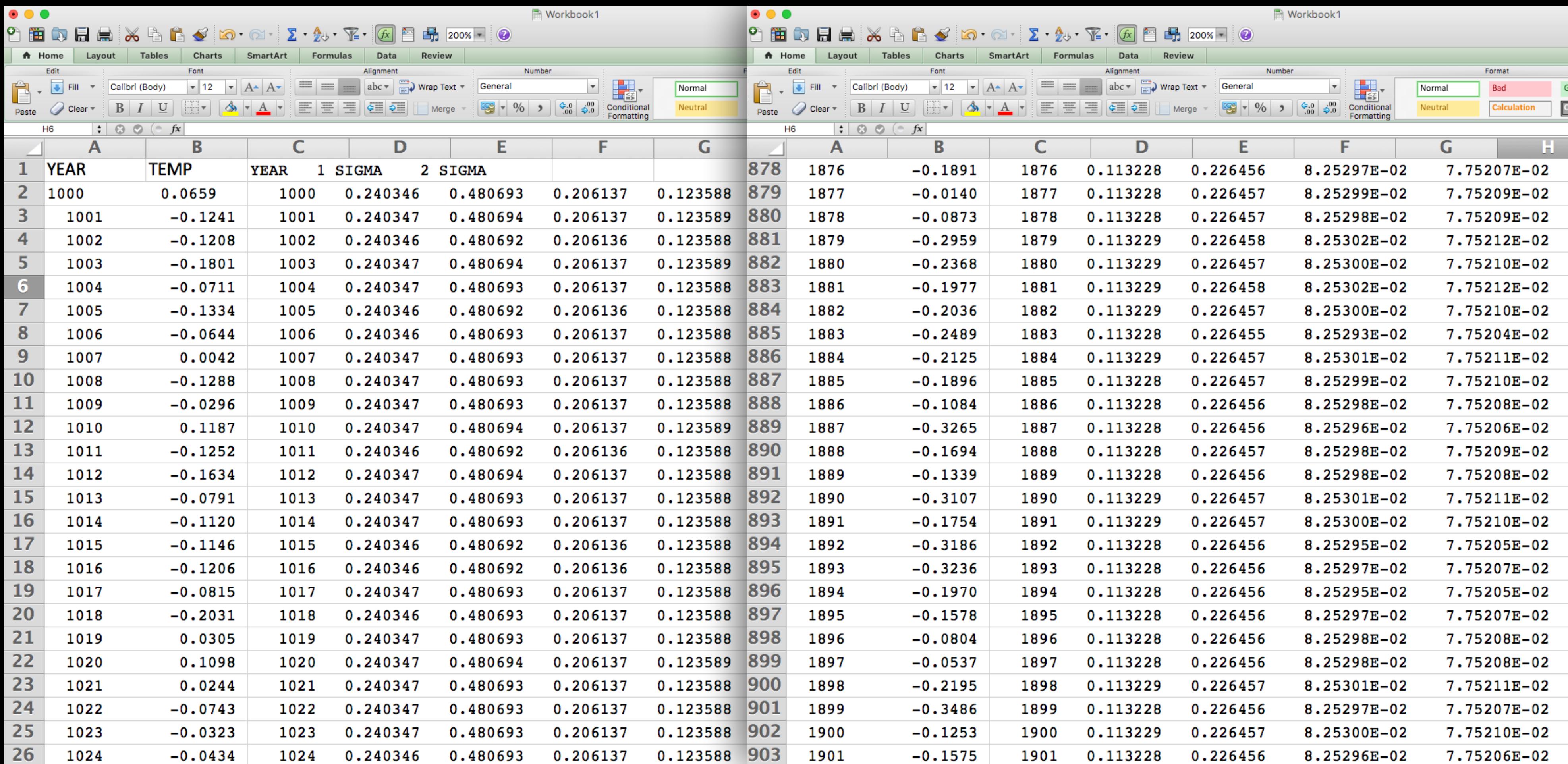
Visualization is a bit like writing: beyond some conventions and constraints regarding symbols, visual grammar, perception, and cognition, visualization **can't be based on “rules” that are set in stone.**

Instead, when designing visualizations, we need to be guided by **reasoned, justifiable choices**.



I. Why should my visualization exist?

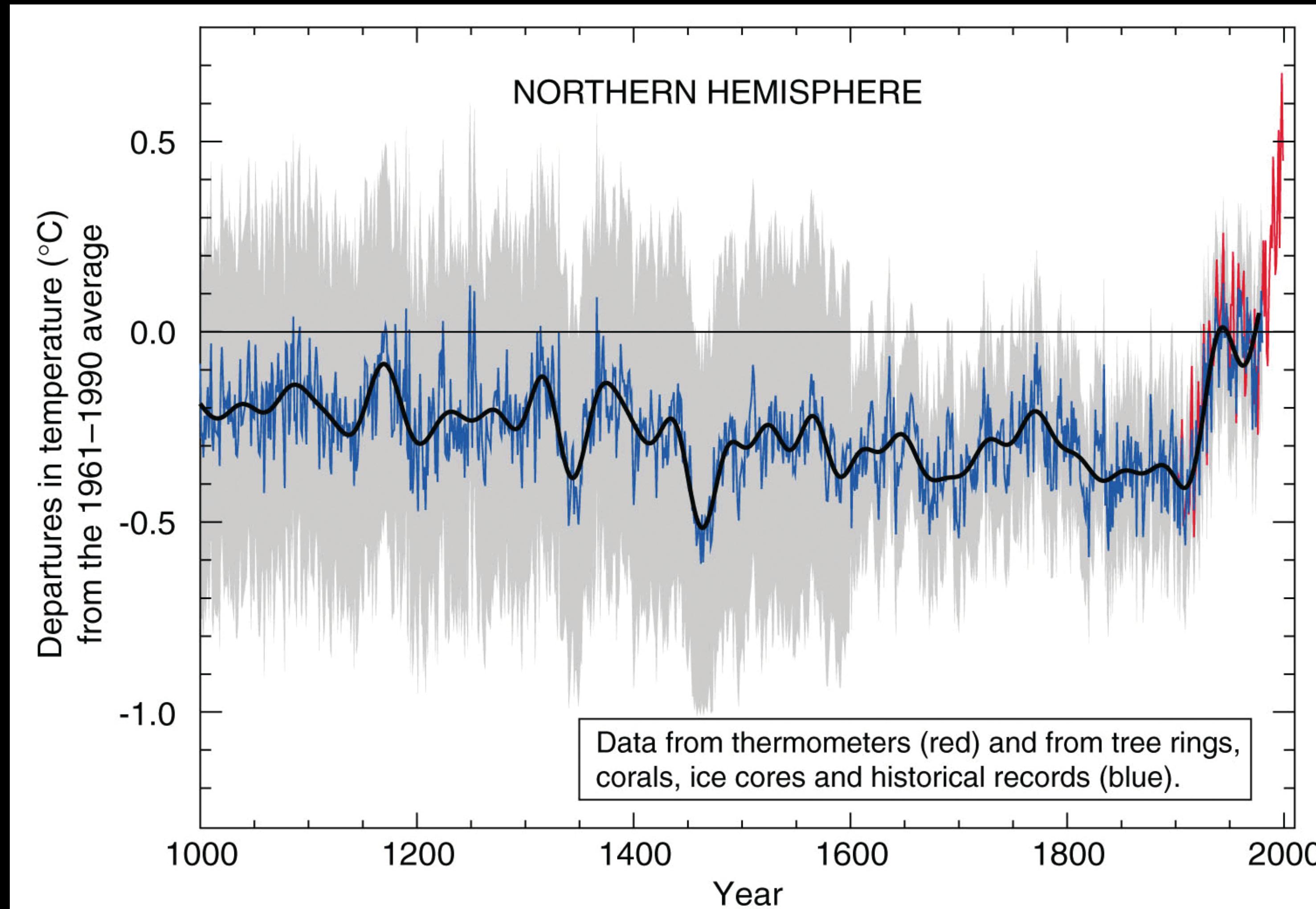
What are tables useful for?



The image displays two side-by-side screenshots of Microsoft Excel, both titled "Workbook1". Each screenshot shows a large data table with approximately 26 rows and 8 columns. The columns are labeled A through H. The first few rows contain header information, while the subsequent rows contain numerical data. The data includes various numbers such as 1000, -0.1241, 0.240346, 0.480693, 0.206137, 0.123588, etc. The second screenshot shows the same data with some cells highlighted in red, indicating errors or specific values of interest.

| | A | B | C | D | E | F | G | | A | B | C | D | E | F | G | H |
|----|------|---------|------|----------|----------|----------|----------|-----|------|---------|------|----------|----------|-------------|-------------|---|
| 1 | YEAR | TEMP | YEAR | 1 SIGMA | 2 SIGMA | | | 878 | 1876 | -0.1891 | 1876 | 0.113228 | 0.226456 | 8.25297E-02 | 7.75207E-02 | |
| 2 | 1000 | 0.0659 | 1000 | 0.240346 | 0.480693 | 0.206137 | 0.123588 | 879 | 1877 | -0.0140 | 1877 | 0.113228 | 0.226457 | 8.25299E-02 | 7.75209E-02 | |
| 3 | 1001 | -0.1241 | 1001 | 0.240347 | 0.480694 | 0.206137 | 0.123589 | 880 | 1878 | -0.0873 | 1878 | 0.113228 | 0.226457 | 8.25298E-02 | 7.75209E-02 | |
| 4 | 1002 | -0.1208 | 1002 | 0.240346 | 0.480692 | 0.206136 | 0.123588 | 881 | 1879 | -0.2959 | 1879 | 0.113229 | 0.226458 | 8.25302E-02 | 7.75212E-02 | |
| 5 | 1003 | -0.1801 | 1003 | 0.240347 | 0.480694 | 0.206137 | 0.123589 | 882 | 1880 | -0.2368 | 1880 | 0.113229 | 0.226457 | 8.25300E-02 | 7.75210E-02 | |
| 6 | 1004 | -0.0711 | 1004 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 883 | 1881 | -0.1977 | 1881 | 0.113229 | 0.226458 | 8.25302E-02 | 7.75212E-02 | |
| 7 | 1005 | -0.1334 | 1005 | 0.240346 | 0.480692 | 0.206136 | 0.123588 | 884 | 1882 | -0.2036 | 1882 | 0.113229 | 0.226457 | 8.25300E-02 | 7.75210E-02 | |
| 8 | 1006 | -0.0644 | 1006 | 0.240346 | 0.480693 | 0.206137 | 0.123588 | 885 | 1883 | -0.2489 | 1883 | 0.113228 | 0.226455 | 8.25293E-02 | 7.75204E-02 | |
| 9 | 1007 | 0.0042 | 1007 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 886 | 1884 | -0.2125 | 1884 | 0.113229 | 0.226457 | 8.25301E-02 | 7.75211E-02 | |
| 10 | 1008 | -0.1288 | 1008 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 887 | 1885 | -0.1896 | 1885 | 0.113228 | 0.226457 | 8.25299E-02 | 7.75210E-02 | |
| 11 | 1009 | -0.0296 | 1009 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 888 | 1886 | -0.1084 | 1886 | 0.113228 | 0.226456 | 8.25298E-02 | 7.75208E-02 | |
| 12 | 1010 | 0.1187 | 1010 | 0.240347 | 0.480694 | 0.206137 | 0.123589 | 889 | 1887 | -0.3265 | 1887 | 0.113228 | 0.226456 | 8.25296E-02 | 7.75206E-02 | |
| 13 | 1011 | -0.1252 | 1011 | 0.240346 | 0.480692 | 0.206136 | 0.123588 | 890 | 1888 | -0.1694 | 1888 | 0.113228 | 0.226457 | 8.25298E-02 | 7.75209E-02 | |
| 14 | 1012 | -0.1634 | 1012 | 0.240347 | 0.480694 | 0.206137 | 0.123588 | 891 | 1889 | -0.1339 | 1889 | 0.113228 | 0.226456 | 8.25298E-02 | 7.75208E-02 | |
| 15 | 1013 | -0.0791 | 1013 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 892 | 1890 | -0.3107 | 1890 | 0.113229 | 0.226457 | 8.25301E-02 | 7.75211E-02 | |
| 16 | 1014 | -0.1120 | 1014 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 893 | 1891 | -0.1754 | 1891 | 0.113229 | 0.226457 | 8.25300E-02 | 7.75210E-02 | |
| 17 | 1015 | -0.1146 | 1015 | 0.240346 | 0.480692 | 0.206136 | 0.123588 | 894 | 1892 | -0.3186 | 1892 | 0.113228 | 0.226456 | 8.25295E-02 | 7.75205E-02 | |
| 18 | 1016 | -0.1206 | 1016 | 0.240346 | 0.480692 | 0.206136 | 0.123588 | 895 | 1893 | -0.3236 | 1893 | 0.113228 | 0.226456 | 8.25297E-02 | 7.75207E-02 | |
| 19 | 1017 | -0.0815 | 1017 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 896 | 1894 | -0.1970 | 1894 | 0.113228 | 0.226456 | 8.25295E-02 | 7.75205E-02 | |
| 20 | 1018 | -0.2031 | 1018 | 0.240346 | 0.480693 | 0.206137 | 0.123588 | 897 | 1895 | -0.1578 | 1895 | 0.113228 | 0.226456 | 8.25297E-02 | 7.75207E-02 | |
| 21 | 1019 | 0.0305 | 1019 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 898 | 1896 | -0.0804 | 1896 | 0.113228 | 0.226456 | 8.25298E-02 | 7.75208E-02 | |
| 22 | 1020 | 0.1098 | 1020 | 0.240347 | 0.480694 | 0.206137 | 0.123589 | 899 | 1897 | -0.0537 | 1897 | 0.113228 | 0.226456 | 8.25298E-02 | 7.75208E-02 | |
| 23 | 1021 | 0.0244 | 1021 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 900 | 1898 | -0.2195 | 1898 | 0.113229 | 0.226457 | 8.25301E-02 | 7.75211E-02 | |
| 24 | 1022 | -0.0743 | 1022 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 901 | 1899 | -0.3486 | 1899 | 0.113228 | 0.226456 | 8.25297E-02 | 7.75207E-02 | |
| 25 | 1023 | -0.0323 | 1023 | 0.240347 | 0.480693 | 0.206137 | 0.123588 | 902 | 1900 | -0.1253 | 1900 | 0.113229 | 0.226457 | 8.25300E-02 | 7.75210E-02 | |
| 26 | 1024 | -0.0434 | 1024 | 0.240346 | 0.480693 | 0.206137 | 0.123588 | 903 | 1901 | -0.1575 | 1901 | 0.113228 | 0.226456 | 8.25296E-02 | 7.75206E-02 | |

Visualization is about patterns, trends, the big picture



Michael E. Mann, Raymond S. Bradley, and Malcolm K. Hughes

Intergovernmental Panel on Climate Change (IPCC), Third Report, 2001



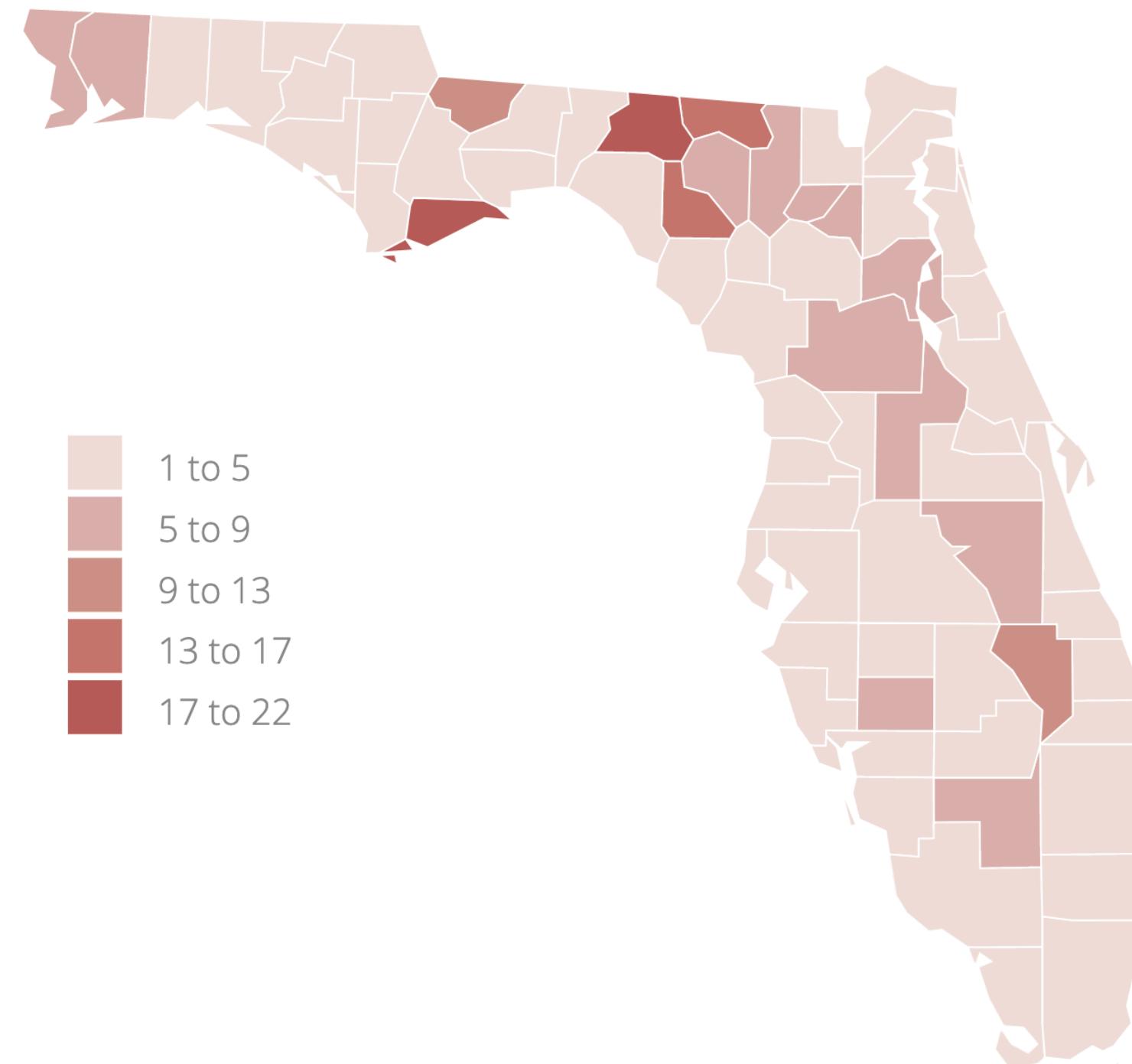
2. What to visualize?

AT SCHOOL WITHOUT A ROOF

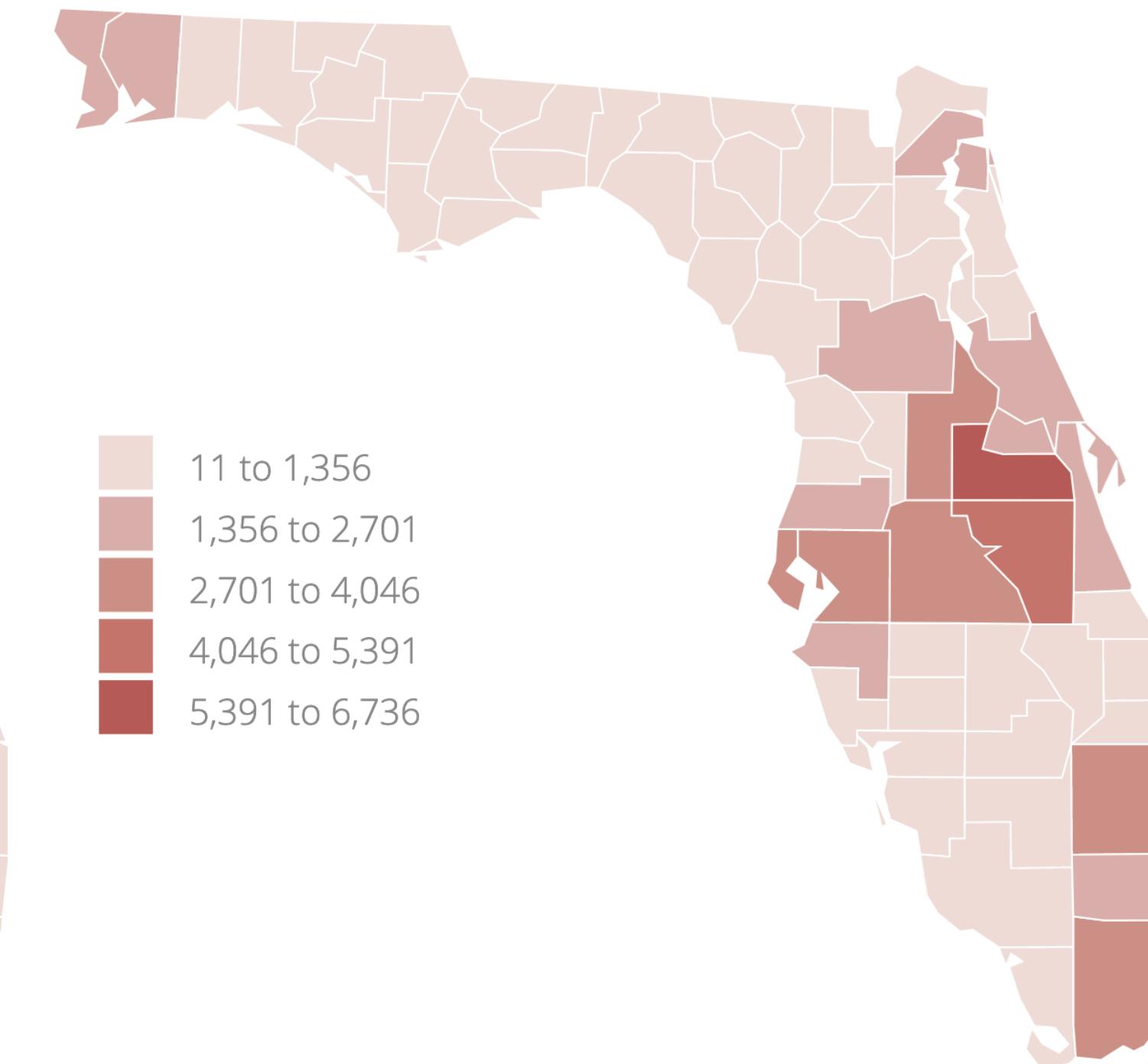
[http://
www.lmelgar.me/
without-a-roof/](http://www.lmelgar.me/without-a-roof/)

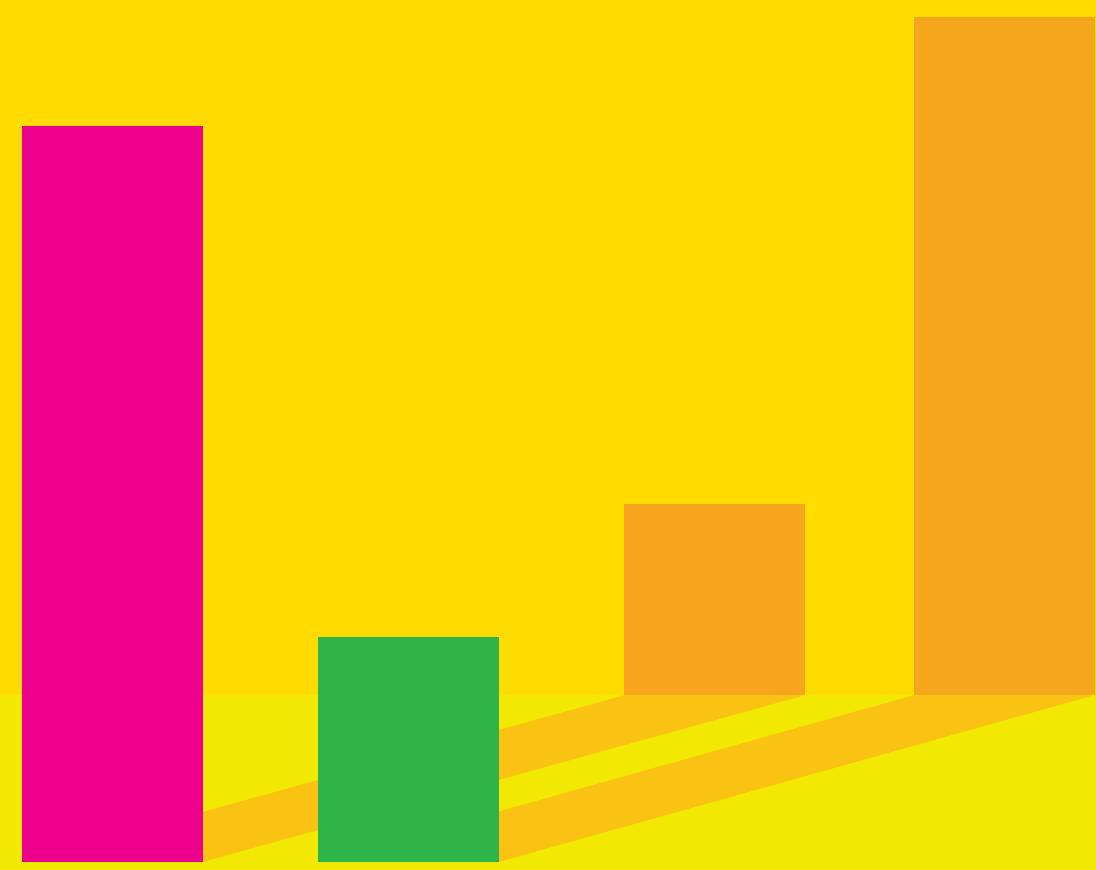
In Florida more than 71,000 students are homeless. During the last decade, this population rocketed as a result of the recession and how hard it has become for the poorest families to find affordable housing.

Percentage Total



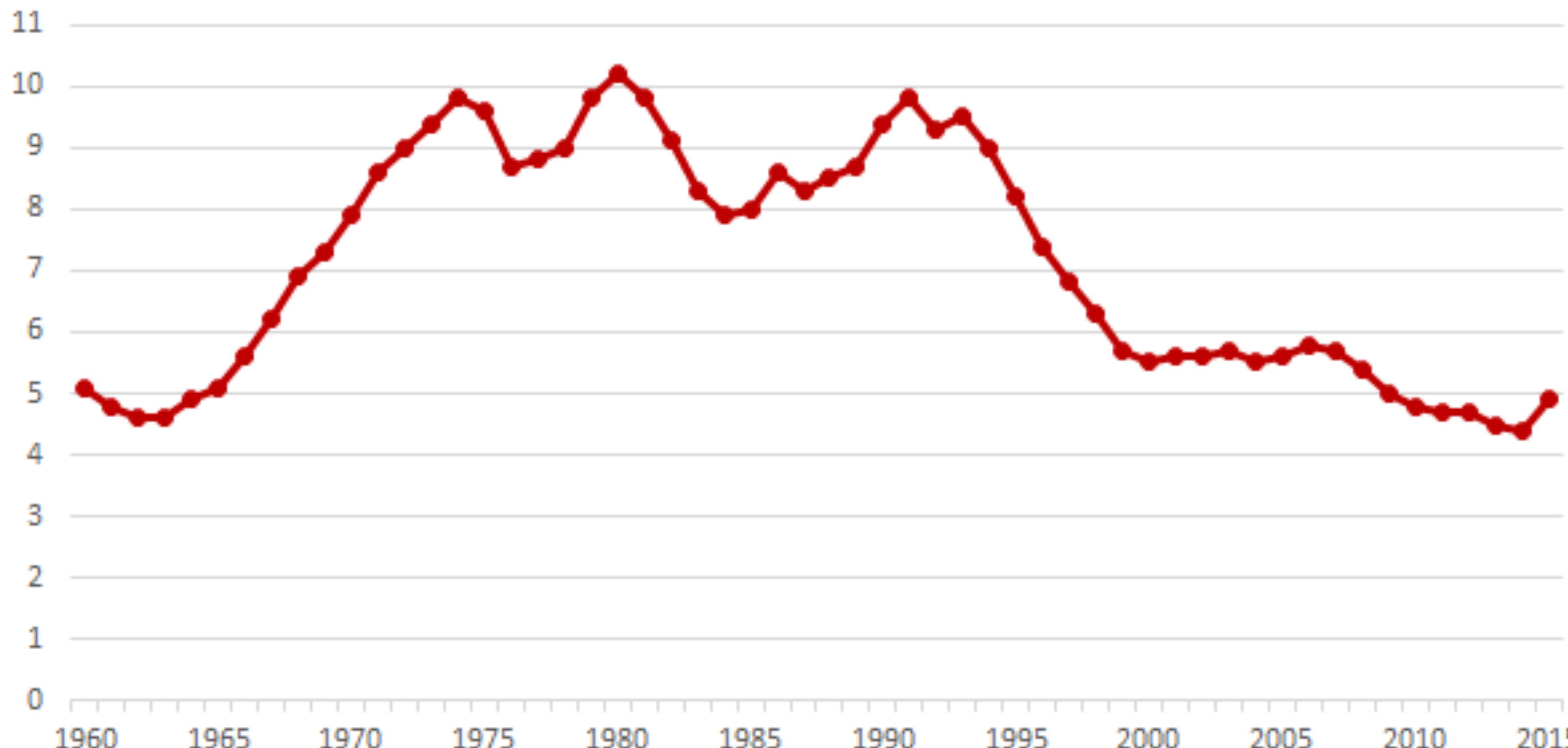
Percentage Total





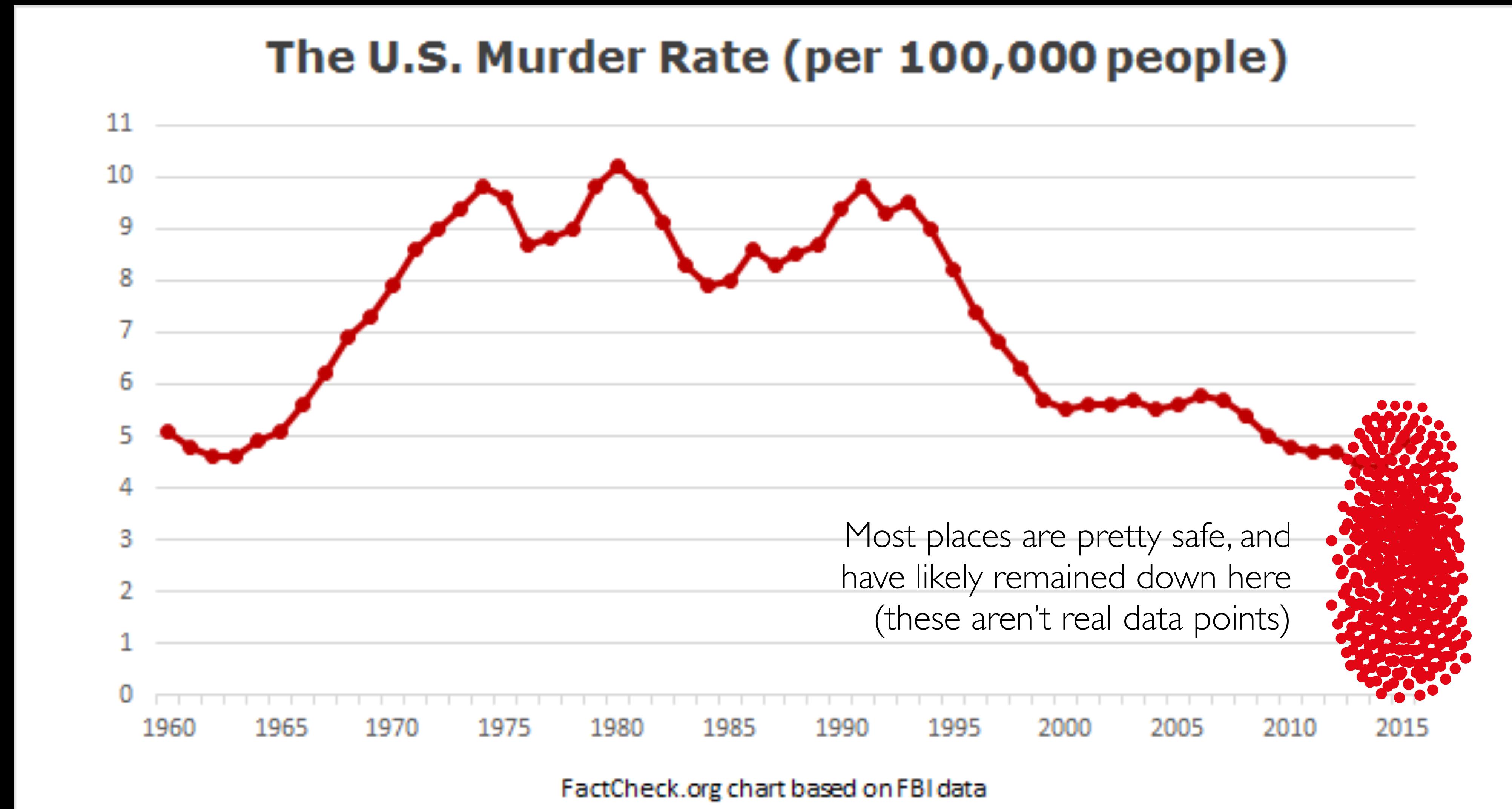
3. How much to visualize?

The U.S. Murder Rate (per 100,000 people)

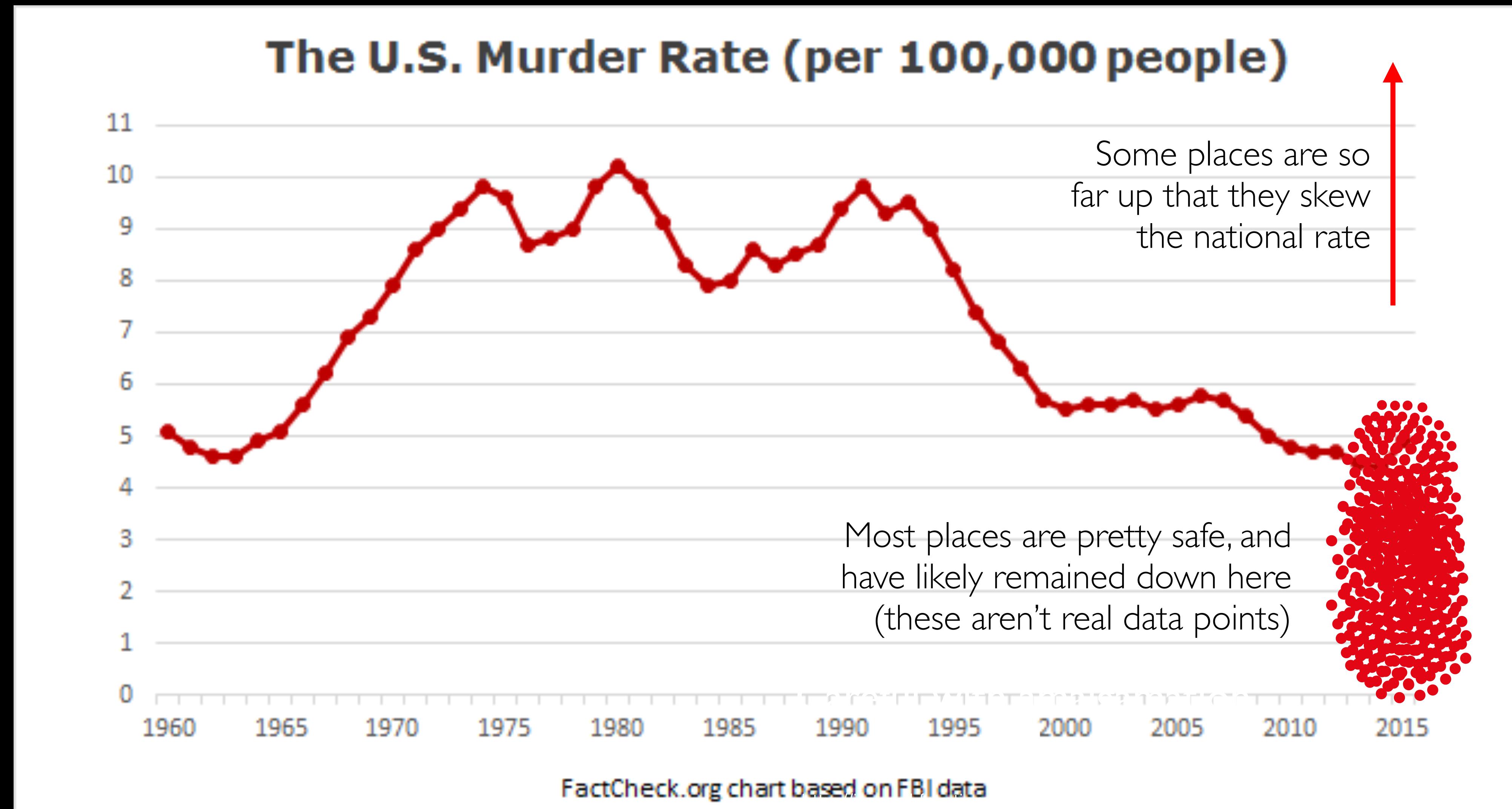


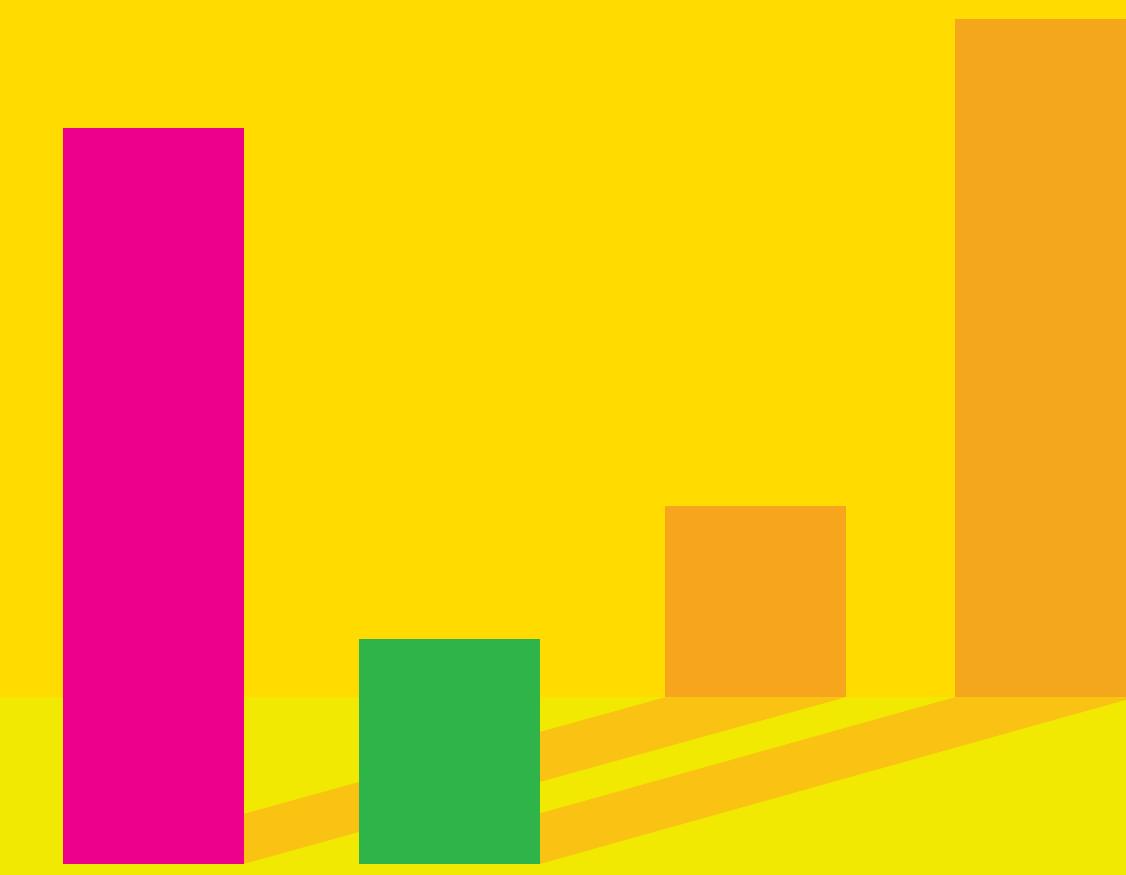
FactCheck.org chart based on FBI data

The danger of aggregating data too much,
and presenting just averages and other statistical summaries



The danger of aggregating data too much,
and presenting just averages and other statistical summaries





4. How to visualize it?

Figure 2 - Main nationalities of arriving migrants – 2016

Greece

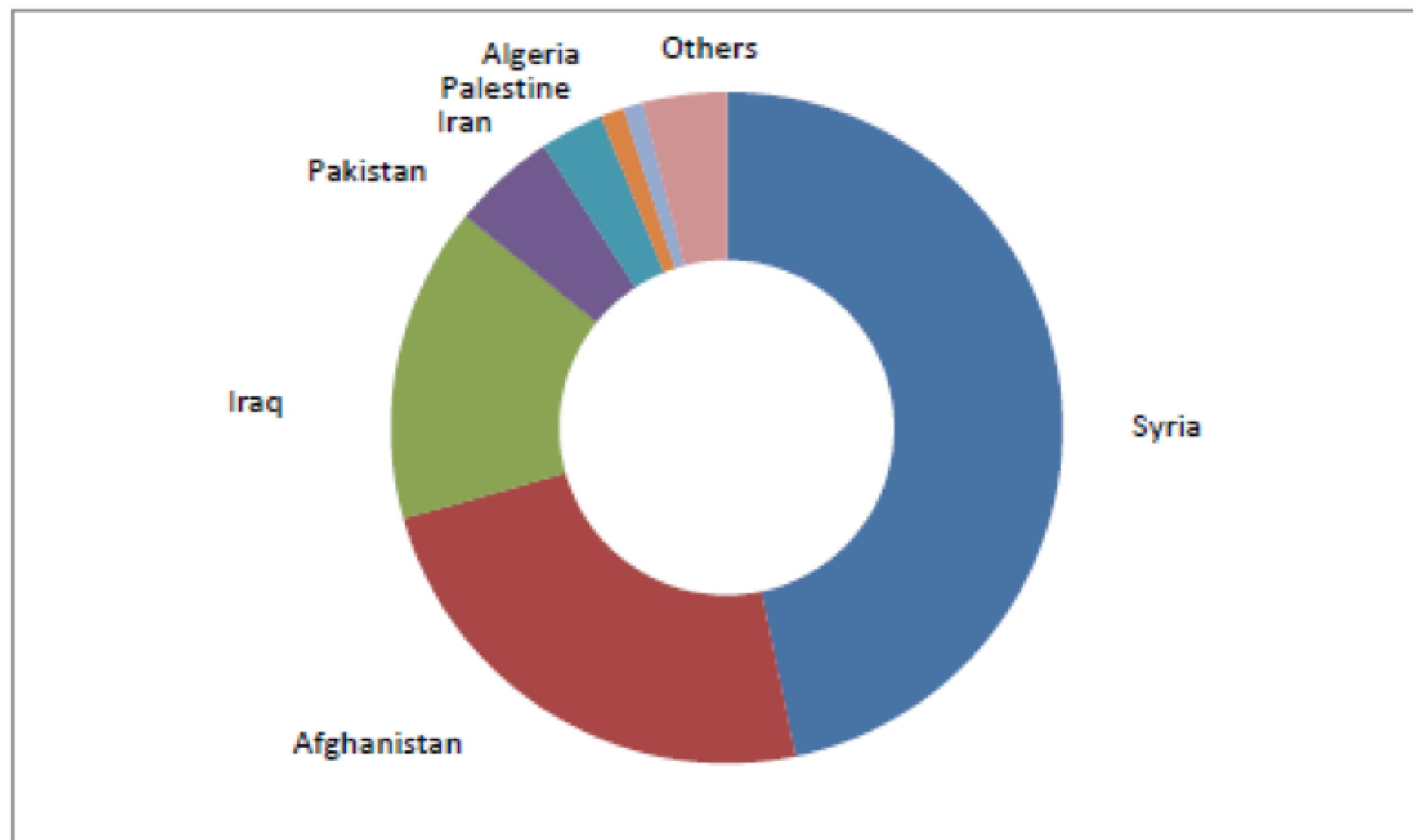
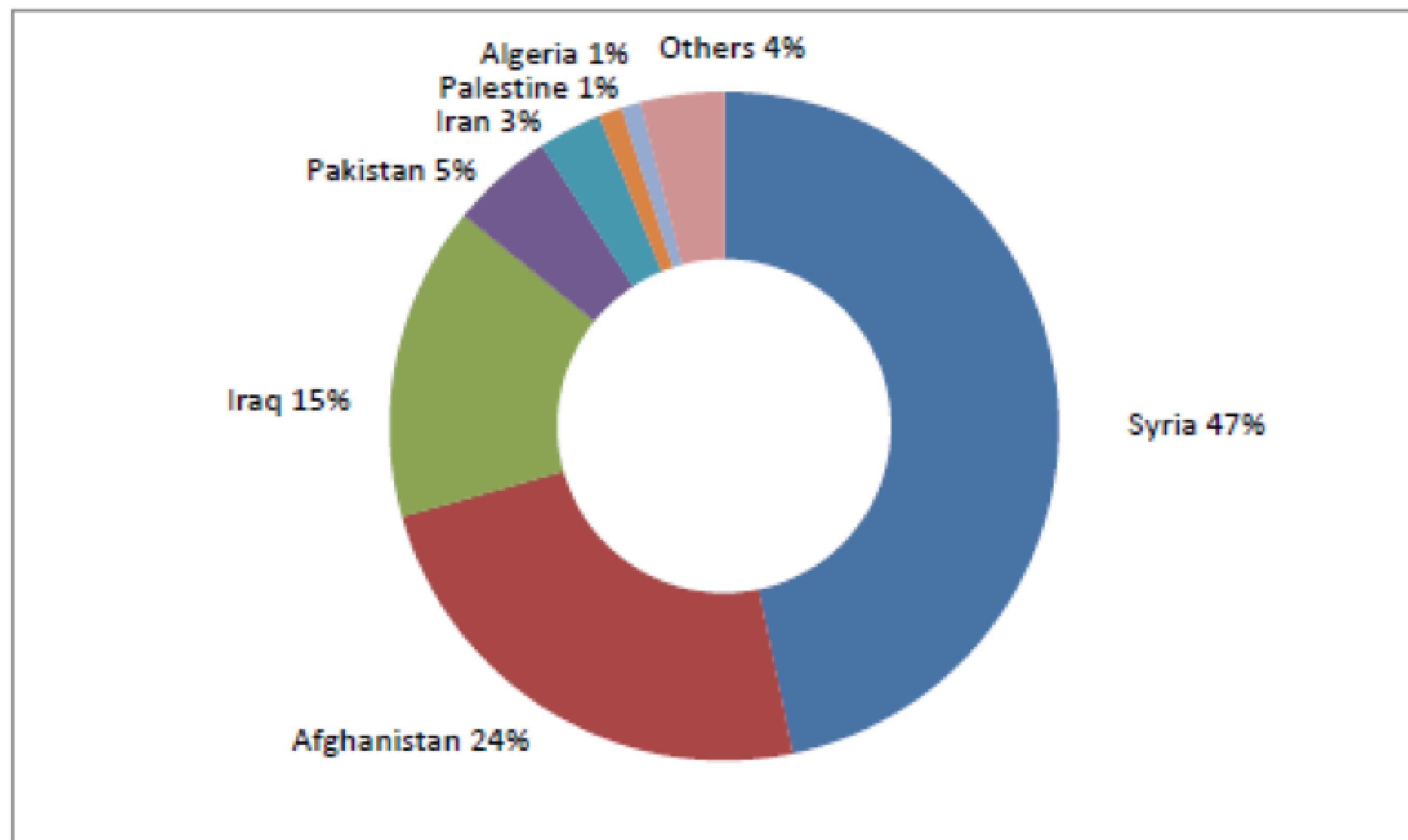
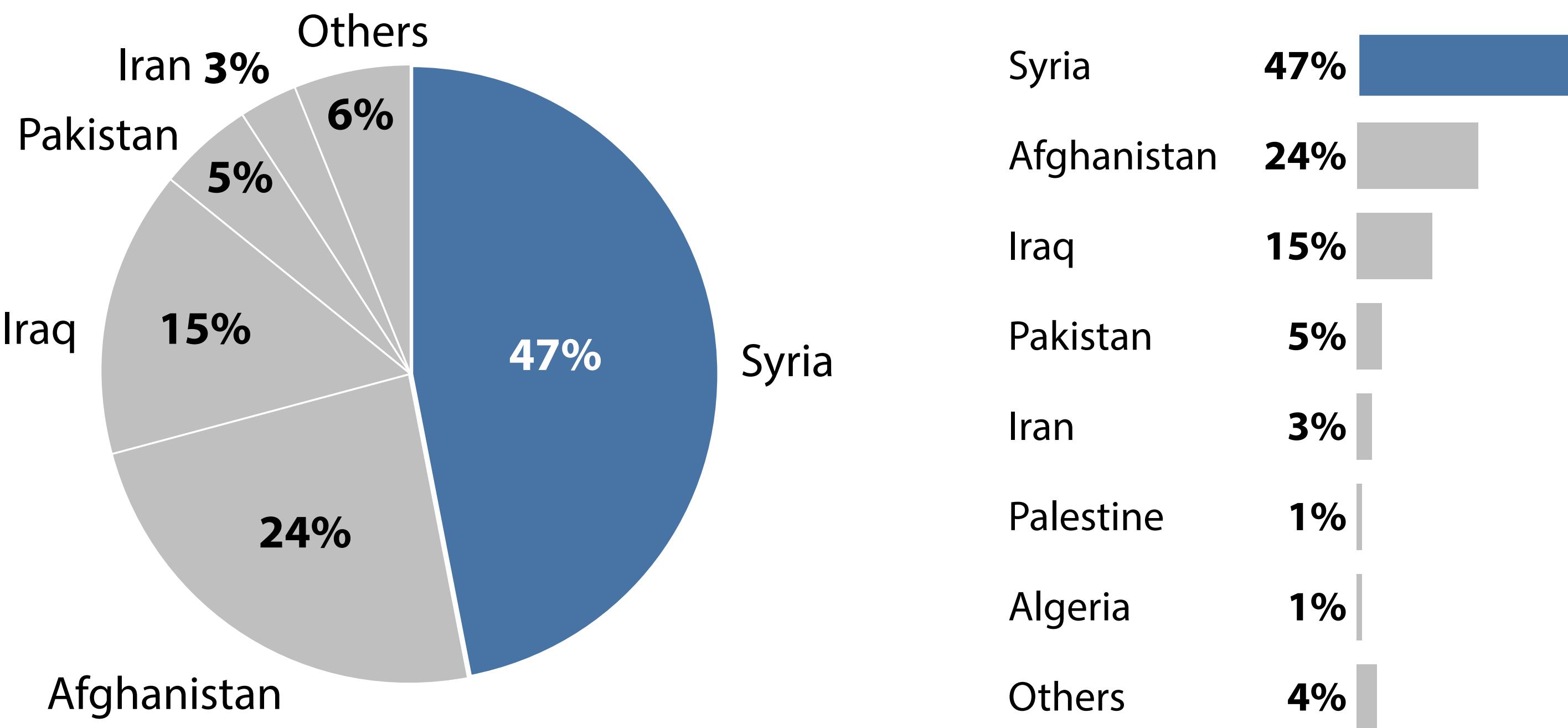
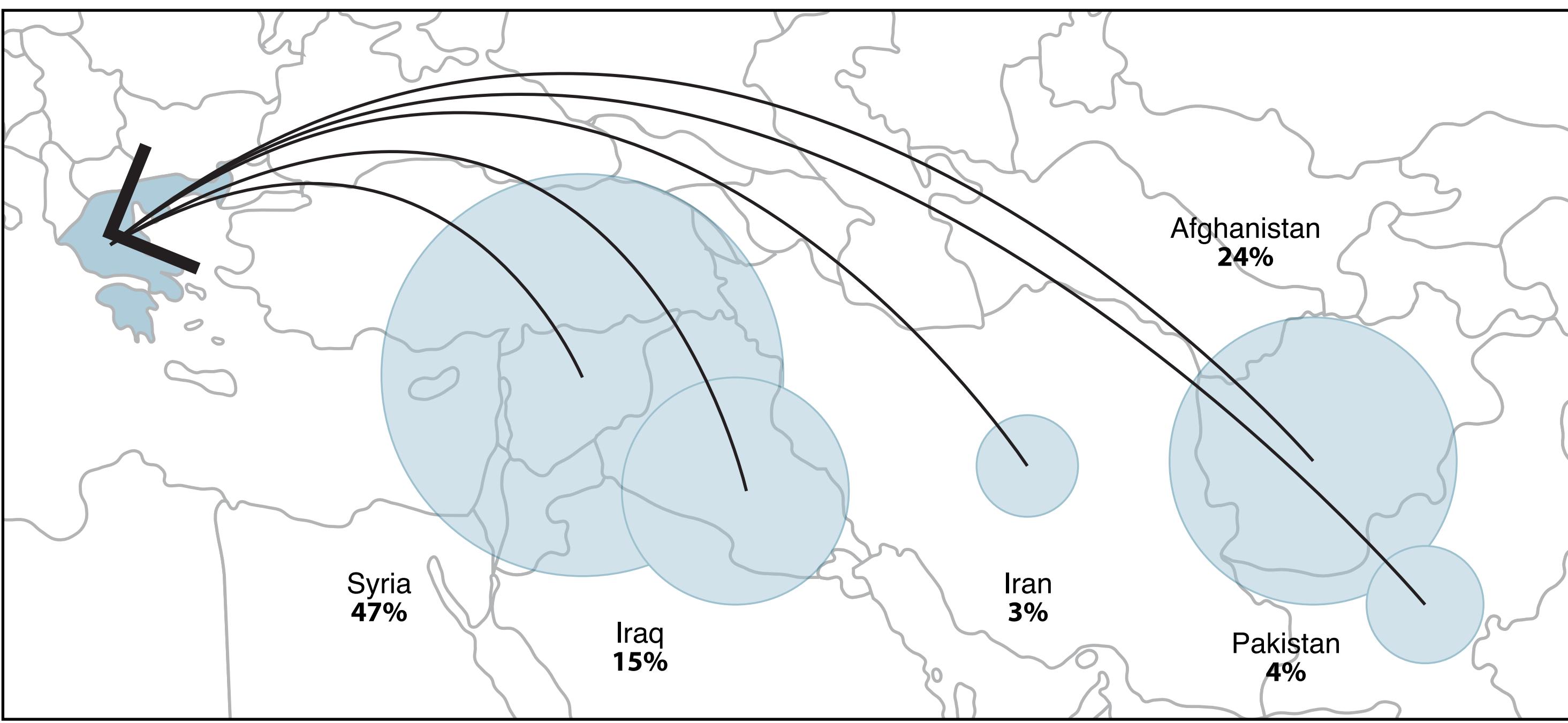


Figure 2 - Main nationalities of arriving migrants – 2016

Greece





The Data Visualisation Catalogue

About • Suggest • Shop • Resources

Search by Function



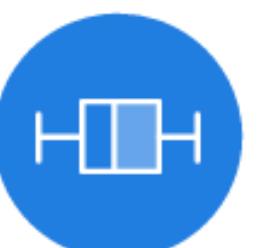
Arc Diagram



Area Graph



Bar Chart



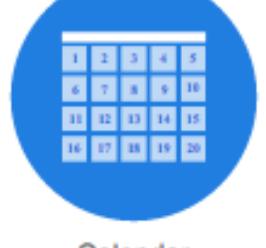
Box & Whisker Plot



Brainstorm



Bubble Chart



Calendar



Chord Diagram



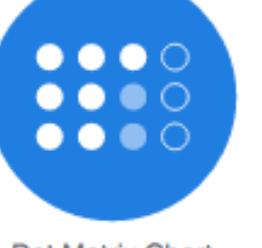
Choropleth Map



Circle Packing



Donut Chart



Dot Matrix Chart



Flow Map



Histogram



Illustration Diagram



Line Graph



Marimekko Chart



Multi-set Bar Chart



Nightingale Rose Chart



Non-ribbon Chord Diagram



Parallel Sets



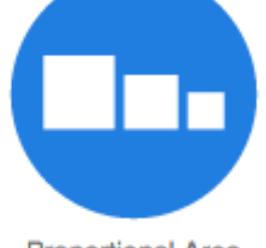
Pictogram Chart



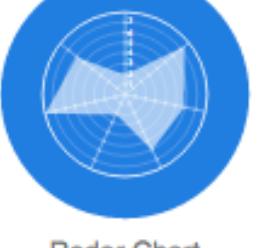
Pie Chart



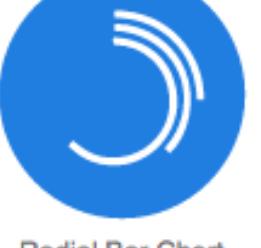
Population Pyramid



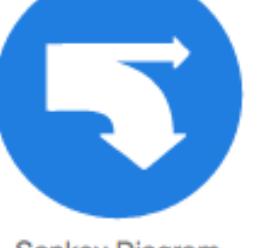
Proportional Area Chart



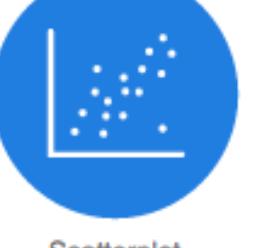
Radar Chart



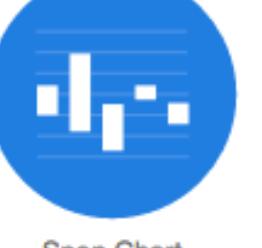
Radial Bar Chart



Sankey Diagram



Scatterplot



Span Chart

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but can also be a target or a long-term average. Can also be used to show sentiment (positive/negative/neutral).

Example FT uses: Trade surplus/deficit, climate change

Correlation

Show the relationship between two or more variables. Be mindful that unless you tell them otherwise, many readers will assume the relationships you show them to be causal (e.g. one causes the other).

Example FT uses: Inflation & unemployment, income & life expectancy

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses: Wealth, deprivation, league tables, constituency election results

Distribution

Show values in a dataset and how often they occur. The shape (or 'view') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses: Income distribution, population (geographic) distribution

Change over Time

Give emphasis to changing trends. These can be short (month-over-month) movements or extended series (decades or centuries). Choosing what time period is important to provide suitable context for the reader.

Example FT uses: Share price movements, economic time series

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses: Fiscal budgets, company structures, national election results

Magnitude

Show size comparisons. These can be relative (e.g. being able to see larger/larger) or absolute (need to see fine differences). Usually these show a ratio or rate, for example, barrels, dollars or people rather than a calculated rate or per cent.

Example FT uses: Commodity production, market capitalisation

Spatial

Used only when precise locations or sequences of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses: Movement of funds, trade, migrants, lawsuits, information: relationship graphs.

Flow

Show the master volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses: Movement of funds, trade, migrants, lawsuits, information: relationship graphs.

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Scatterplot

The standard way to show the relationship between two continuous variables, each with its own axis.

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to reflect the 'shape' of the data.

Line

The standard way to show a changing time series. If data are irregular, consider using dots to represent data points.

Stacked column

A simple way of showing part-to-whole relationships but can be difficult to read with many components.

Bar

See above. Good when the data are not time series and labels have long category names.

Proportional stacked bar

A good way of showing the size and proportion of data at the same time - as long as the data are not too complicated.

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets, which can include +/- components.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (e.g. disagreement/agree).

Line + column

A good way of showing the relationship between an amount (columns) and a rate (line).

Ordered column

See above.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data.

Violin plot

Similar to a box plot but can show even more complex distributions (data that cannot be summarised with a simple average).

Line + column

Good for showing change over time - but usually best with one series of data at a time.

Column

Columns work well for showing change over time - but usually best with one rate (column) and a rate (line).

Pie

As per standard pie charts, good for multiple series. Can become tricky to read with more than 2 series.

Paired column

See above.

Dot

As per standard dot plots, good for multiple series - but it's difficult to accurately compare the size of the segments.

Stock price

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Dot strip plot

Similar to a pie chart but the centre can be a good way of making space to include more information about the data (e.g. totals).

Dot plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Slope

Good for showing how values have changed over time or vary between categories.

Lollipop chart

Lollipops draw more attention to the data value than standard bar charts, but can also show rank and value effectively.

Cumulative curve

A good way of showing how unequal a distribution is - x axis is cumulative frequency, y axis is always a measure.

Connected scatterplot

A good way of showing changing data for two variables where there is a repeating pattern of progression.

Barcode plot

Like dot strip plots, good for displaying all the data in a single row, good for highlighting individual values.

Fan chart (projections)

Like dot strip plots, good for showing the uncertainty in future projections - usually this grows the further forward to projections.

Area chart

Use with care - these are good at showing changes to totals, but sometimes changes across multiple categories can be very difficult.

Barcode plot

Like dot strip plots, good for displaying all the data in a single row, good for highlighting individual values.

Fan chart (projections)

Like dot strip plots, good for showing the uncertainty in future projections - usually this grows the further forward to projections.

Veronoi

A way of mapping points into areas - any point within each area is closer to the central point than any other centroid.

Proportional stacked bar

Use for hierarchical don't-care relationships can be difficult to read when there are many small segments.

Sunburst

Another way of visualising hierarchical part-to-whole relationships. Use sparingly if at all for obvious reasons.

Isotype (pictogram)

Use when there are some instances - use only with whole numbers as it does not slice off an arm to represent a decimal.

Arc

A hemispherical, often used for visualising political results in parliaments.

Lollipop chart

Lollipop charts draw more attention to the data value than standard bar charts, but can also show rank and value effectively.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly). The presence of shading precision in quantity.

Gridplot

Good for showing % information, they work best when used with discrete categories and work well in multiple layout form.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Venn

Generally only used for schematic representation.

Circle timeline

Good for showing discrete values of varying size across multiple categories (e.g. participants by continent).

Selogram

Another alternative to the circle timeline where there are big variations in the data.

Radar chart

A space-efficient way of showing value of multiple variables - but the shape is not always as clear as a circle, and it does not HAVE to start at zero (but preferable).

Heat map

Grid-based data values mapped with an intensity colour scale. As per standard heatmap, but not snapped to an admin/political unit.

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Ben Bern; Li Fausto; Graham Parish; Billy Ehrenberg; Paul McCullagh; Martin Stalbe
Inspired by the Graphic Contourine by Jon Schwabish and Severine Ribecca

ft.com/vocabulary

<http://www.datavizcatalogue.com/>

<https://github.com/ft-interactive/chart-doctor/blob/master/visual-vocabulary/Visual-vocabulary.pdf>

Alberto Cairo • University of Miami • www.thefunctionalart.com • Twitter: @albertocairo