Econ 2148, spring 2019
Data visualization

Maximilian Kasy

Department of Economics, Harvard University

Agenda

- ▶ One way to think about statistics: Mapping data-sets into numerical summaries that are interpretable by readers.
- ▶ Estimates, tests, confidence sets, predictions ...
- ▶ We can also map data-sets into visual representations.
- ▶ How to think systematically about these mappings?
- ▶ How to implement them?
- ▶ What are good design practices?

Takeaways for this part of class

- ▶ The "layered grammar of graphics" provides a framework for describing mappings from data to visual representations.
- ▶ It allows to systematically implement visualizations, and to come up with new types of visualizations.
- ▶ This grammar is the foundation for *ggplot2*, a popular graphics package for R.
- ▶ Good design practices for visualization:
    1. Show the data.
    2. Reduce the clutter.
    3. Integrate the text and the graph.

Why discuss a "grammar of graphics?"

Wickham (2010):

- ▶ It gives us a framework to think about graphics.
- ▶ It shortens the distance from mind to paper.
- ▶ It allows to iteratively update a plot, changing a single feature at a time.
- ▶ It encourages the use of customized graphics, rather than relying on generic named graphics.
- ▶ It helps to discover new types of graphics.
- ▶ It helps to understand how *ggplot2* works.

Components of the "layered grammar of graphics"

1. A data-set and set of mappings from variables to aesthetics.
2. One or more layers, with each layer having
   - one geometric object,
   - one statistical transformation.
3. One scale for each aesthetic mapping used.
4. A coordinate system.
5. The facet specification.

Aesthetics

- ▶ x-position.
- ▶ y-position.
- ▶ Color.
- ▶ Shape.
- ▶ Size / thickness.

Statistical transformations

- ▶ Identity.
- ▶ Bin counts.
- ▶ Statistics for box plots.
- ▶ Contour lines.
- ▶ 1d density estimate.
- ▶ Quantile regression.
- ▶ Smoothed conditional mean.
- ▶ Removing duplicates.
- ▶ ...

Geometric objects and Scales

- **Geometric objects**:
  - 0 dimensional: Point, text.
  - 1 dimensional: Path, line.
  - 2 dimensional: Polygon, interval.
- **Scales**: Mapping from data to aesthetic attributes.
  - Inverse of scale: Guide.
  - Allows reader to map visualization back to data.
  - E.g., legends, axes.

Coordinate systems and faceting

- ▶ **Coordinate system**:
  Map the position of objects onto the plane of the plot.
  - ▶ Cartesian.
  - ▶ Logarithmic.
  - ▶ Polar.
  - ▶ Projection (from higher dimensions).
- ▶ **Faceting**: Create small multiples.
  - ▶ Divide the data based on some variable.
  - ▶ Create analogous plots for each subset.

Some examples

Practice problem

For each of the following examples from Healy (2018),

1. discuss it in terms of the "layered grammar of graphics",
2. predict what the resulting plot will be.

```
p = ggplot(data = gapminder,
           mapping = aes(x = gdpPercap,
                         y=lifeExp))
p + geom_smooth()
```

geom_smooth()
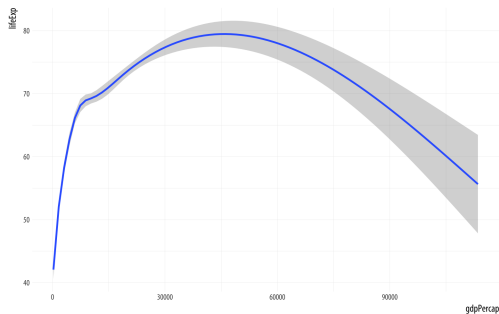
Next slide:

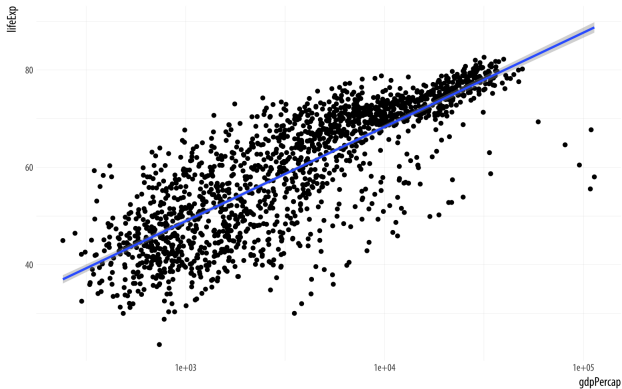```
p + geom_point() + geom_smooth()
```

# geom_point() + geom_smooth()
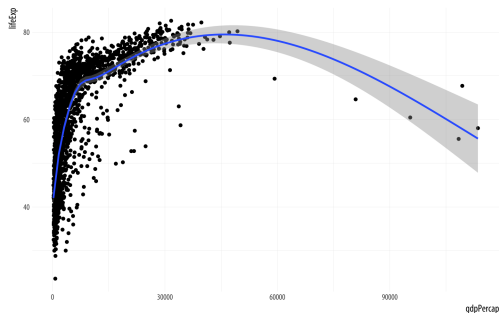


Next slide:

```
p + geom_point() +
    geom_smooth() +
    scale_x_log10()
```

Log scale

Next slide:

```
p + geom_point(alpha = 0.3) + geom_smooth() +
    scale_x_log10(labels = scales::dollar) +
    labs(x = "GDP Per Capita",
         y = "Life Expectancy in Years",
         title = "Economic Growth and Life Expectancy",
         subtitle = "Data points are country-years",
         caption = "Source: Gapminder.")
```

Labeled plot

Economic Growth and Life Expectancy

Data points are country-years

Life Expectancy in Years

80

60

40

$1,000 $10,000 $100,000
GDP Per Capita
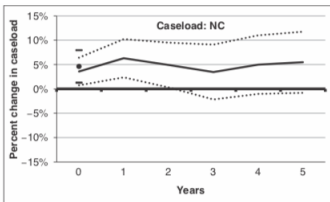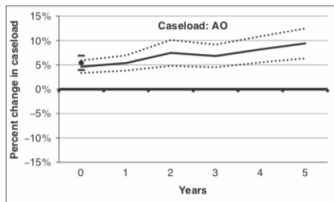
Source: Gapminder.
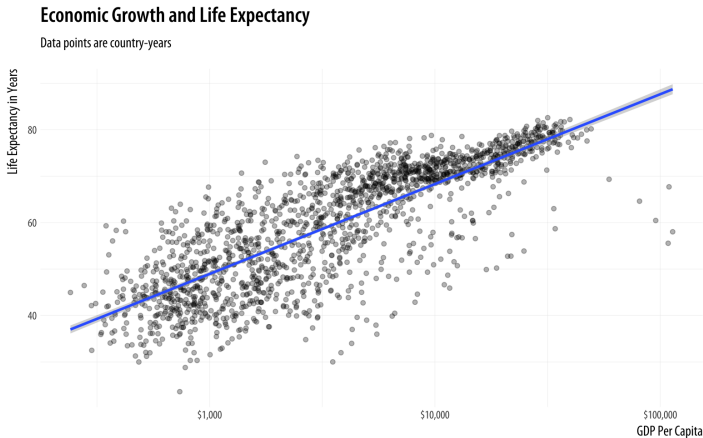
Good practices of data visualization

- ► Schwabish (2014):
    1. Show the data.
    2. Reduce the clutter.
    3. Integrate the text and the graph.
- ► We will go through a series of graphs, discuss their problems, and a possible improved version.

Practice problem

For each of the following "before" graphs, discuss how they are violating the proposed "good practices."

Before
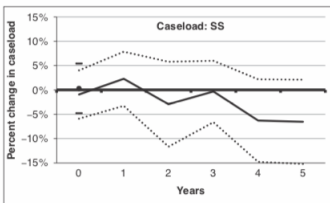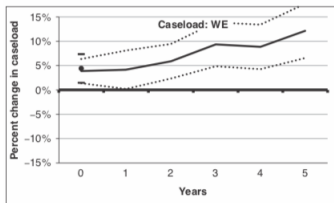
Problems

- ▶ A graph should emphasize the data, but
  - ▶ the darkest and thickest line is the 0 percent grid line,
  - ▶ rather than the coefficient line and the standard errors.
- ▶ Unneeded clutter: y-axis labels, percentage signs, tick marks.
- ▶ What do AO, NC, WE, and SS mean?
- ▶ Proposed improvements:
  - ▶ The darkest line shows the coefficient estimate,
  - ▶ the grid lines are lightened.
  - ▶ 2 sets of axis labels are eliminated
  - ▶ as are the % signs,
  - ▶ repeated title is moved to common title.

After

Implied Impulse Response Functions for Different Caseloads
(Percent change)

Before
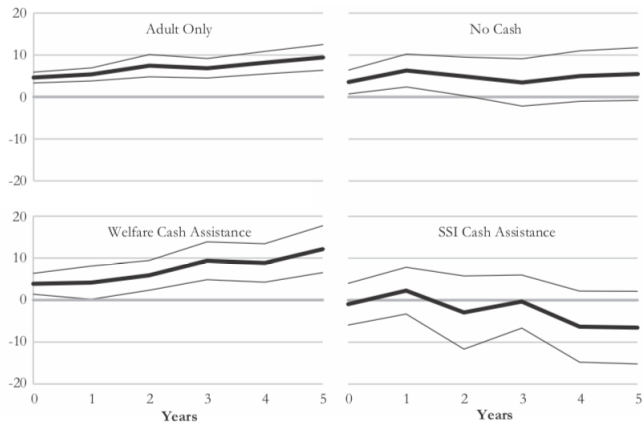
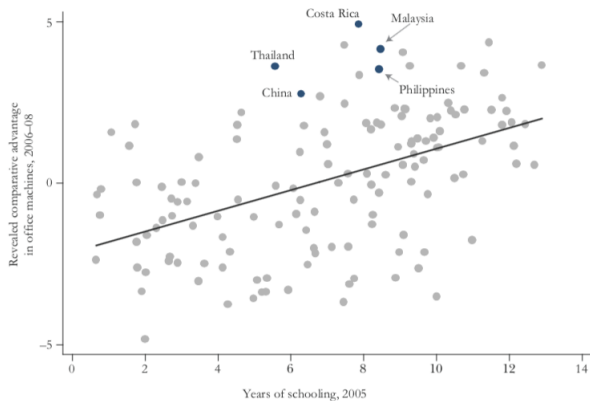

Education and Exports of Office Machines

Problems

- ▶ Hard to find specific countries in the haystack of labels and dots.
- ▶ Proposed improvements:
  - ▶ Eliminate all labels other than for the 5 countries discussed in text.
  - ▶ Spell out country names.
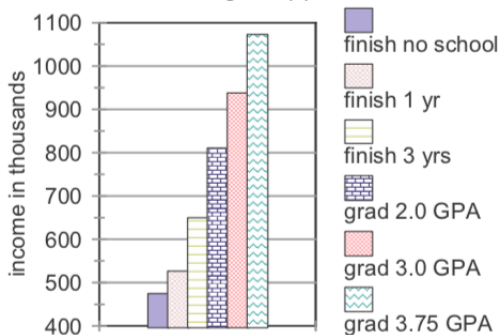 - ▶ Make these 5 data points darker, the rest lighter.

After

Before

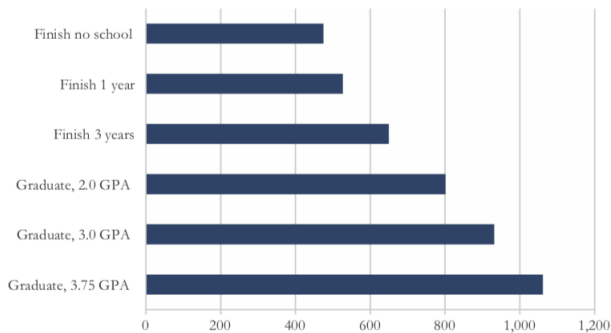**Figure 2 Discounted Expected Lifetime Earnings, VN(t')**

Problems

- ▶ Column chart does not start at zero.
- ▶ Different colors for each bar, which is not necessary.
- ▶ Proposed improvements:
  - ▶ Axis starting at zero.
  - ▶ Rotate figure horizontally,
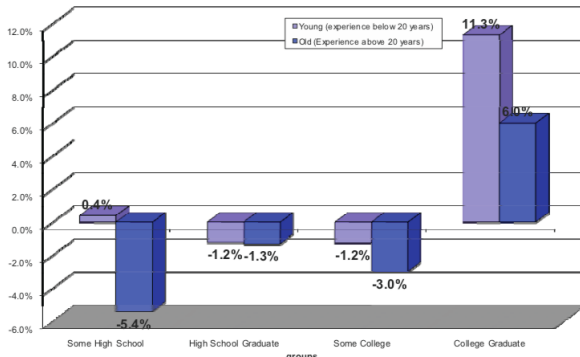 - ▶ which makes room for labels that are integrated with the chart.

After

Discounted Expected Lifetime Earnings, VN(t')
(Income in thousands)

Before

Change in real weekly wages of US-born workers by group, 1990-2006
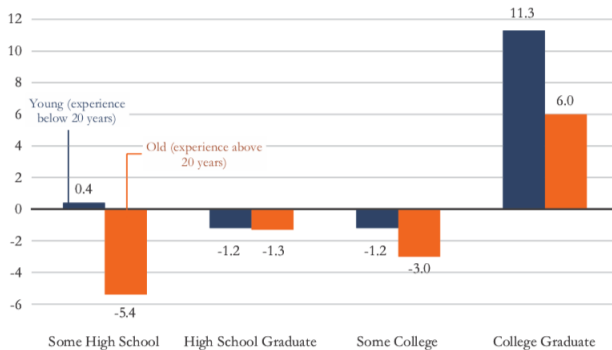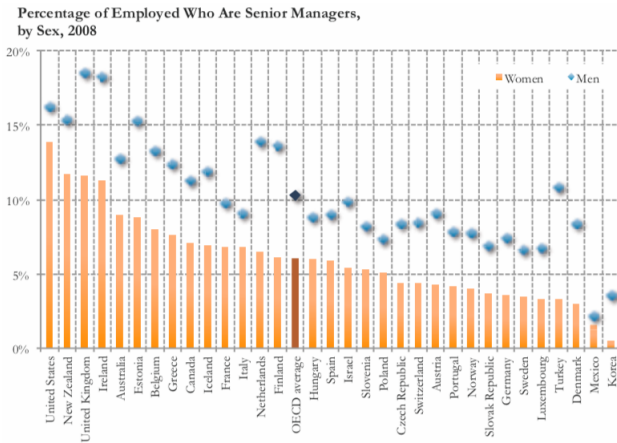
Problems

- ▶ The third dimension does not plot data values,
- ▶ but it does add clutter and can distort the information.
- ▶ Proposed improvements:
  - ▶ Cancel the 3D treatment.
  - ▶ Integrate the disconnected legend with the graph.
  - ▶ Insert the common baseline to permit a more effective comparison among groups.

After

Change in real weekly wages of US-born workers by group, 1990–2006
(Percent)

Before

Percentage of Employed Who Are Senior Managers, by Sex, 2008

Problems

▶ The same kinds of data are plotted using different types of encoding.
  ▶ It is difficult to compare location (diamonds) with length (bars).
  ▶ The bars take up much more space than the diamonds.
  ▶ The points are far away from the columns, with no visual connection.
▶ The columns are darker at the bottom than at the top, where the data are encoded.
▶ Heavy grid lines, redundant percent signs, the labels are vertical.
▶ Proposed improvements:
  ▶ Data encoded similarly for men and women.
  ▶ Title, units, and legend integrated and placed at the top-left.
  ▶ Country labels rotated horizontally and incorporated in chart.
  ▶ Connecting lines to help with comparison.
  ▶ The average value for the OECD as a whole is an unfilled circle.

After

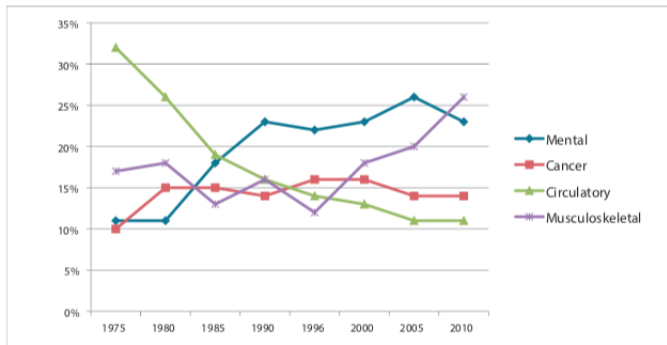Percentage of Employed Who Are Senior Managers, by Gender, 2008
(percent)
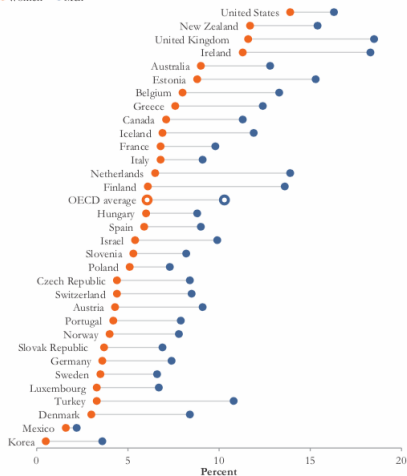
Before

27. Initial DI Worker Awards by Major Cause of Disability—Calendar Years 1975-2010
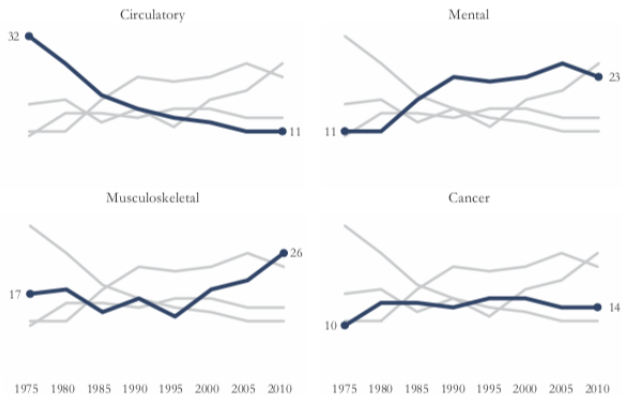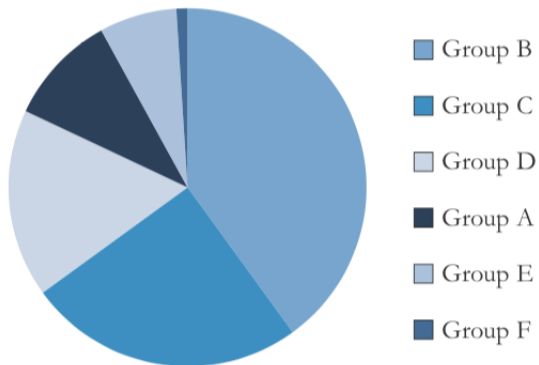
Problems

- ▶ Spaghetti chart:
  To many lines imply any single trend will be obscured.

- ▶ Data markers on every point make it hard to follow any single series.

- ▶ The legend is far from the data,
  the order of the legend does not match the order of the lines.

- ▶ Proposed improvements:
  - ▶ Create smaller charts in series ("sparklines" or "small multiples").
  - ▶ Contrast between light and dark to highlight specific trends.
  - ▶ Label at either end of the main line in each set,
    instead of y-axes.

After

Initial DI Worker Awards by Major Cause of Disability—
Calendar Years 1975–2010
(Percent)

Before

Problems

- ▶ Pie charts force readers to make comparisons
  using the areas or the angles,
  which our visual perception does not accurately support.
  (Donuts are even worse.)
- ▶ Proposed improvements:
  - ▶ Bar chart: best suited for comparing different segments.
    (though less efficient for part-to-whole comparisons)
  - ▶ Plus signs at the bottom to emphasize that the columns sum to
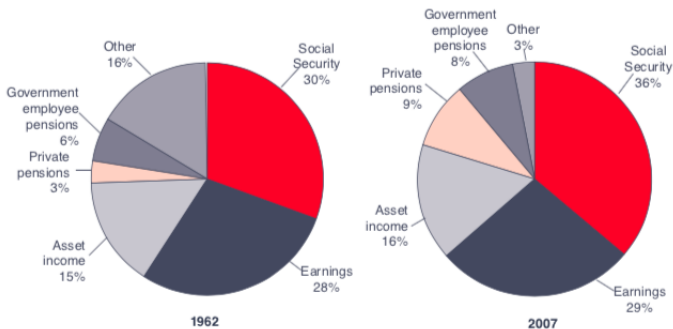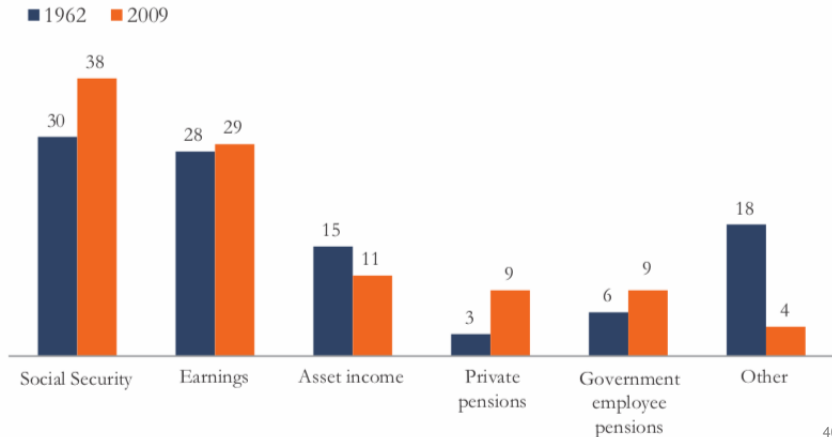   100%.

After
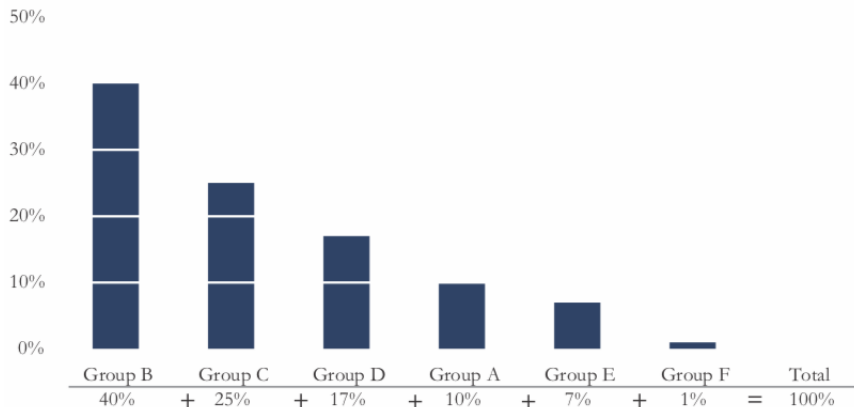
Before

Problems

- ▶ Same problems as for previous example.
- ▶ Several alternatives for proposed improvement:
  1. Paired column chart.
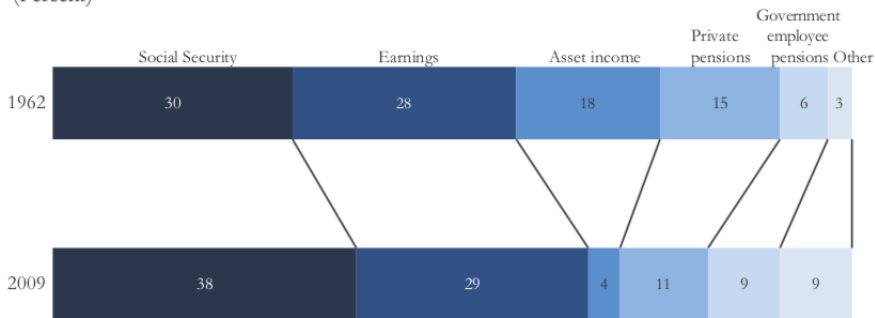 2. Stacked bar chart.
  3. Slope chart.

After, paired column chart

Shares of Aggregate Income, 1962 and 2009
(Percent)

■ 1962   ■ 2009

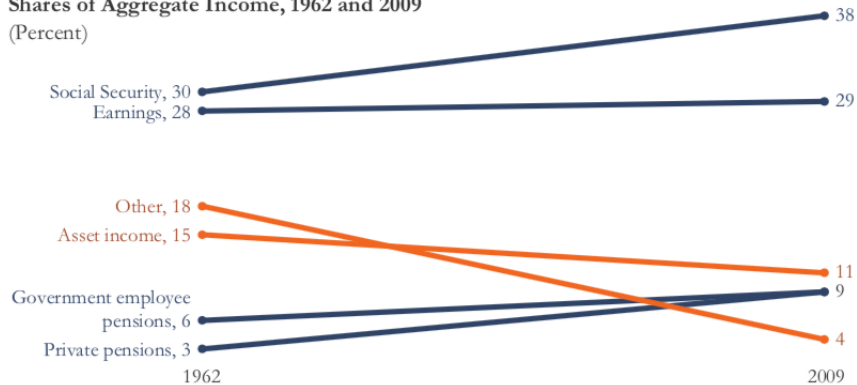|	Social Security	Earnings	Asset income	Private pensions	Government employee pensions	Other
1962	30	28	15	3	6	18
2009	38	29	11	9	9	4

After, stacked bar chart

Shares of Aggregate Income, 1962 and 2009
(Percent)

After, slope chart

Shares of Aggregate Income, 1962 and 2009 (Percent)

References

- *Wickham, H. (2010). A layered grammar of graphics.* Journal of Computational and Graphical Statistics, *19(1):3–28.*
- *Schwabish, J. A. (2014). An economist's guide to visualizing data.* Journal of Economic Perspectives, *28(1):209–34.*
- *Healy, K. (2018).* Data Visualization: A Practical Introduction. *Princeton University Press.*