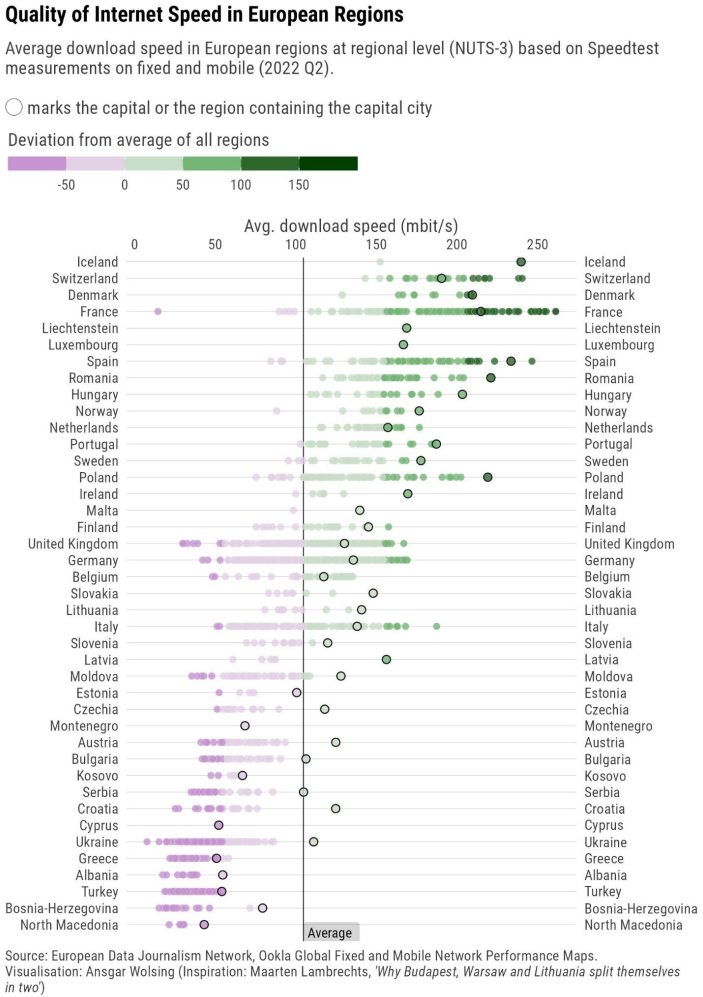
Oct 7, 2024

Data Visualization

Week 3. Visualizing amounts

Reminder

- variable types
- visual aesthetics
- plot analysis (variable + type + aesthetics)
- coordinates
- colors



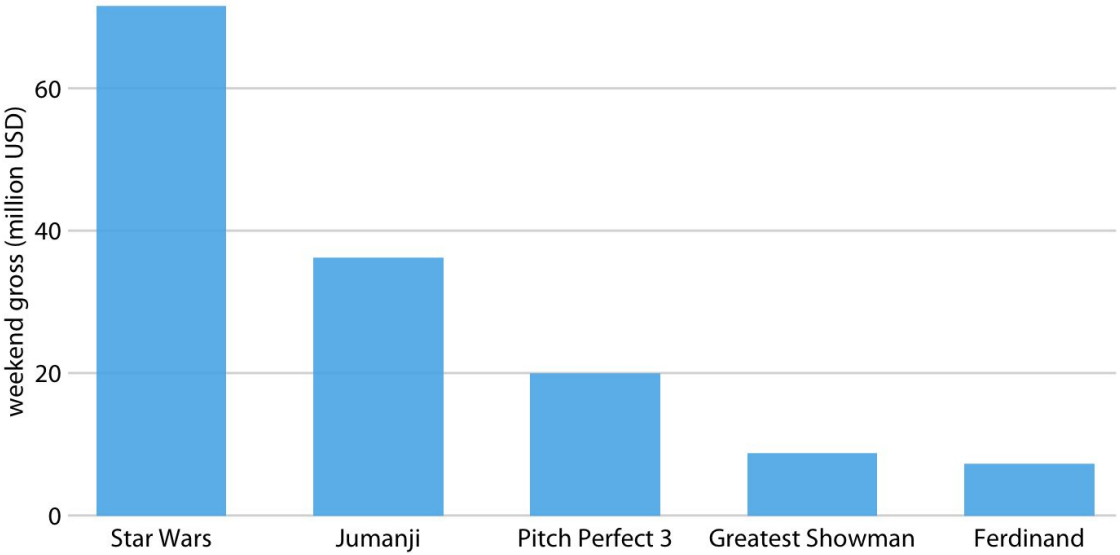
Introduction

In visualization of amounts:

- Bar plots,
- Grouped bar plots,
- Stacked bar plots can be used.

As an alternative to bar plots, dot plots and heatmaps can also be employed.

1. Barplots



= a numeric + a categorical variable

Problems in barplots #1

One of the issues is the random ordering of bars when there is no logical arrangement among the labels.
How it can be solved?

Problems in barplots #1

One of the issues is the random ordering of bars when there is no logical arrangement among the labels.
To make the plot easier to read, the bars should be arranged in ascending or descending order.

Problems in barplots #2

Another issue is that the label names on the axes may take up too much horizontal space or even overlap, making them unreadable.
How it can be solved?

Problems in barplots #2

Another issue is that the label names on the axes may take up too much horizontal space or even overlap, making them unreadable.
In such cases, the most effective solution is to flip the axes.

2. Grouped Barplots

When visualizing amounts, if the dataset contains multiple categorical variables, grouped bar charts can be used.

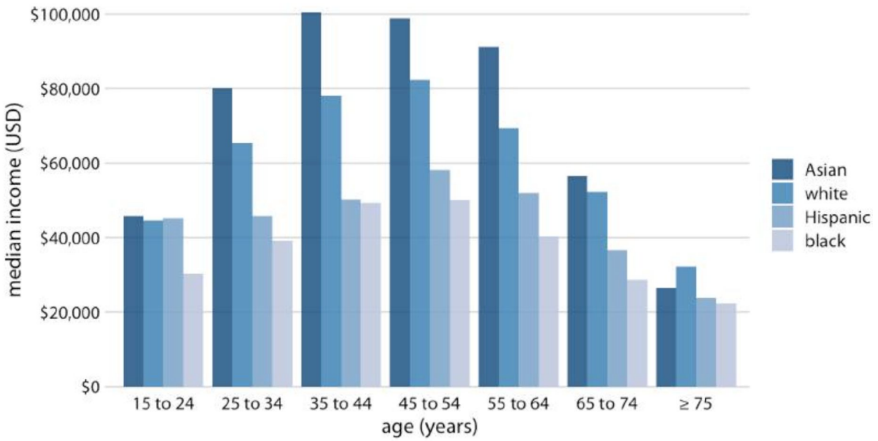


Figure 6-7. 2016 median US annual household income versus age group and race. Age groups are shown along the x axis, and for each age group there are four bars, corresponding to the median income of Asian, white, Hispanic, and black people, respectively. Data source: US Census Bureau.

2. Grouped Barplots

Although the previous plot was prepared correctly, it is a difficult chart to read when comparing the races.

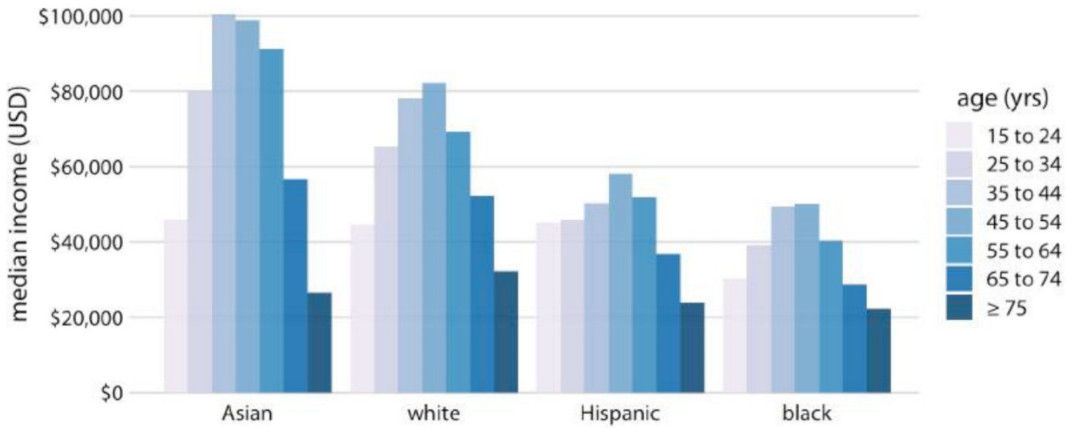


Figure 6-8. 2016 median US annual household income versus age group and race. In contrast to Figure 6-7, now race is shown along the x axis, and for each race we show seven bars according to the seven age groups. Data source: US Census Bureau.

2. Grouped Barplots

In such cases, for convenience, a four-panel barplot can be created instead of a grouped bar chart.

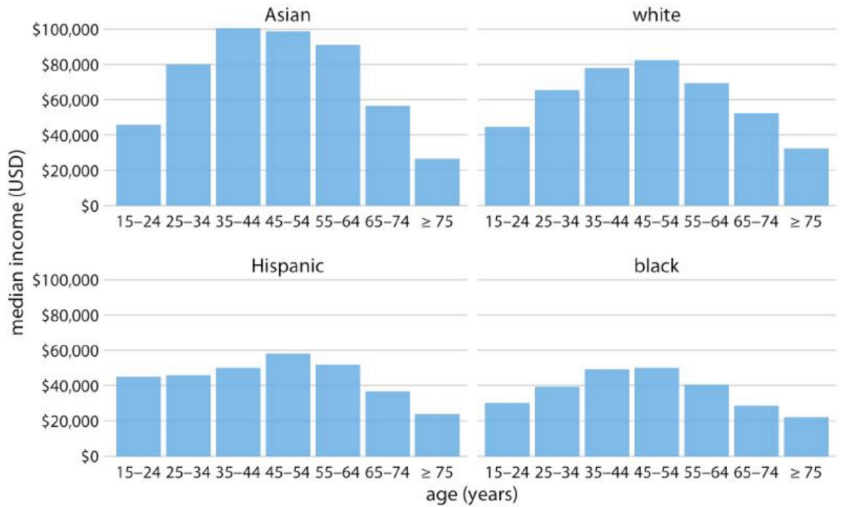


Figure 6-9. 2016 median US annual household income versus age group and race. Instead of displaying this data as a grouped bar plot, as in Figures 6-7 and 6-8, we now show the data as four separate regular bar plots. This choice has the advantage that we don't need to encode either categorical variable by bar color. Data source: US Census Bureau.

3. Stacked Barplots

Grouped barplots are not useful when we want to compare totals. In this case, stacked barplots can be used.

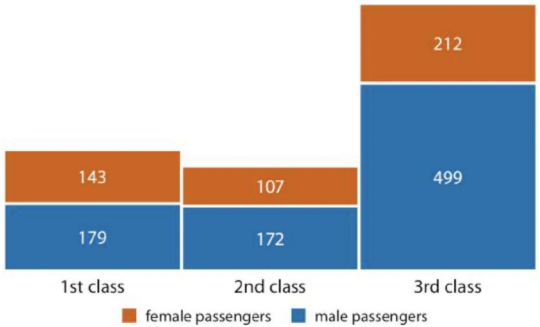


Figure 6-10. Numbers of female and male passengers on the Titanic *traveling in 1st, 2nd, and 3rd class. Data source: Encyclopedia Titanica.*

However, when the number of levels of the categorical variable is 4 or more in such plots, reading the plot becomes difficult. Therefore, **it should be preferred when the number of levels is 2 or 3.**

4. Dotplots

The most significant drawback of bar charts is that they need to start from a zero baseline to accurately represent the quantities they depict. As the number of levels and the amounts increase, this can make the charts harder to read. In such cases, dot plots provide a good alternative.

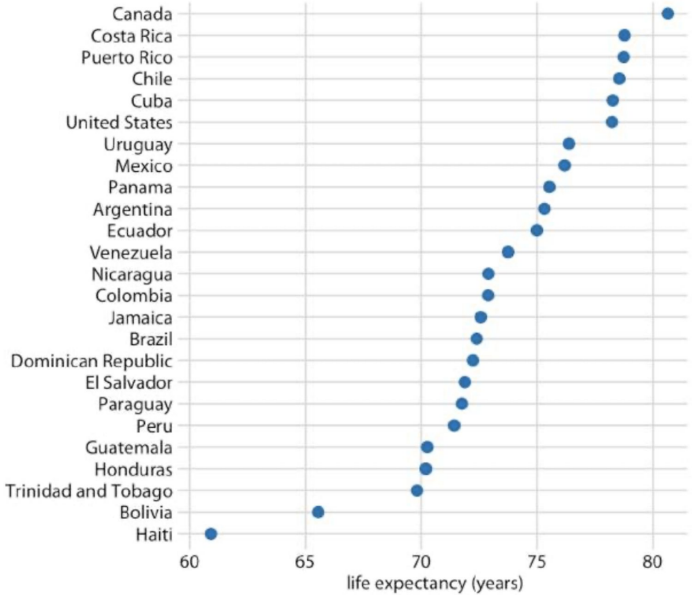


Figure 6-11. Life expectancies of countries in the Americas, for the year 2007. Data source: Gapminder.

4. Dotplots

What if we insisted on using a bar chart instead of a dot plot?

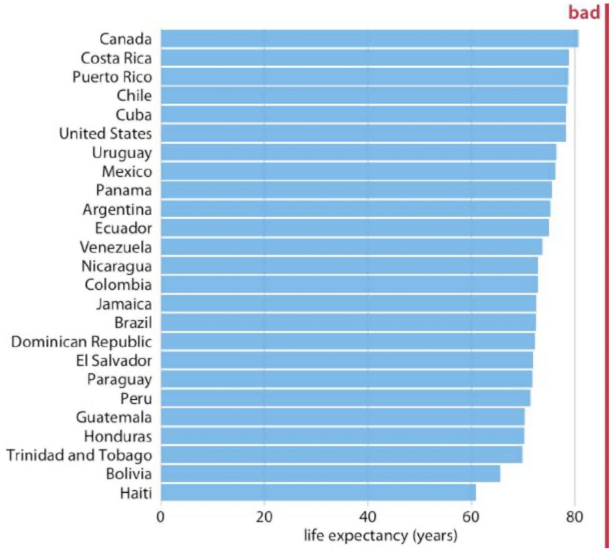


Figure 6-12. Life expectancies of countries in the Americas, for the year 2007, shown as bars. This dataset is not suitable for being visualized with bars. The bars are too long and they draw attention away from the key feature of the data, the differences in life expectancy among the different countries. Data source: Gapminder.

5. Heatmaps

When there are two categorical variables in the dataset, grouped or stacked bar charts can be used. However, as the number of levels of the categorical variables increases, these charts become harder to read.

In such cases, a heat map is a useful alternative.

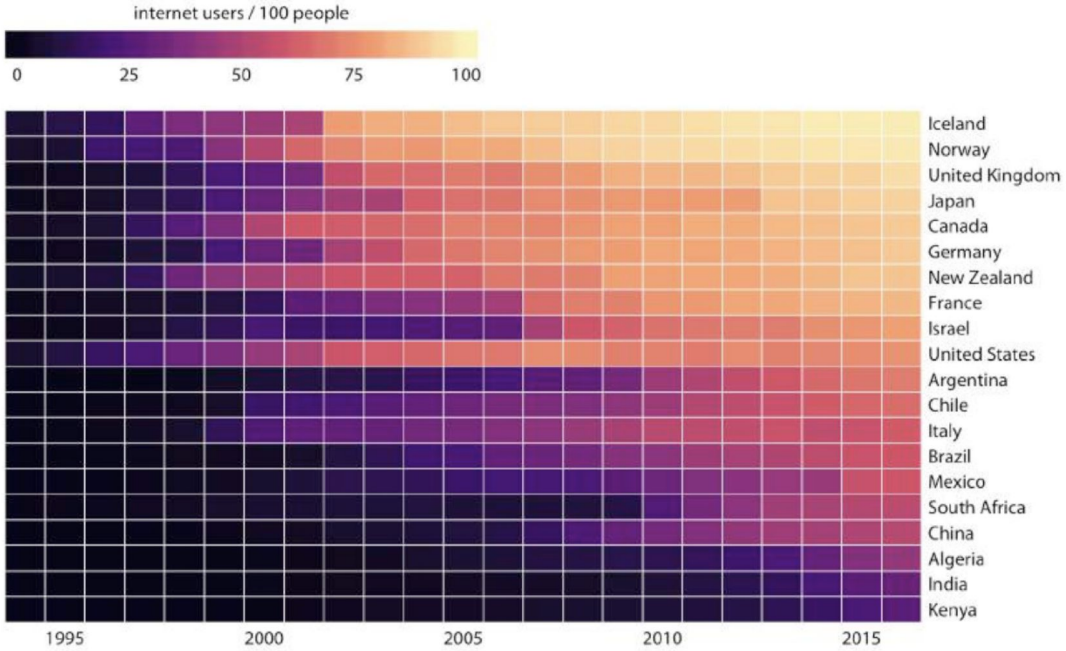


Figure 6-14. Internet adoption over time, for select countries. Color represents the percent of internet users for the respective country and year. Countries were ordered by percent internet users in 2016. Data source: World Bank.

5. Heatmaps

If there is no logical relationship among the levels of the categorical variable in the dataset (if they are not ordered), it is important for ease of readability to sort the levels in increasing or decreasing order based on the values of the corresponding continuous variable. **In a heat map, this arrangement should be made according to the transition of color intensities.**

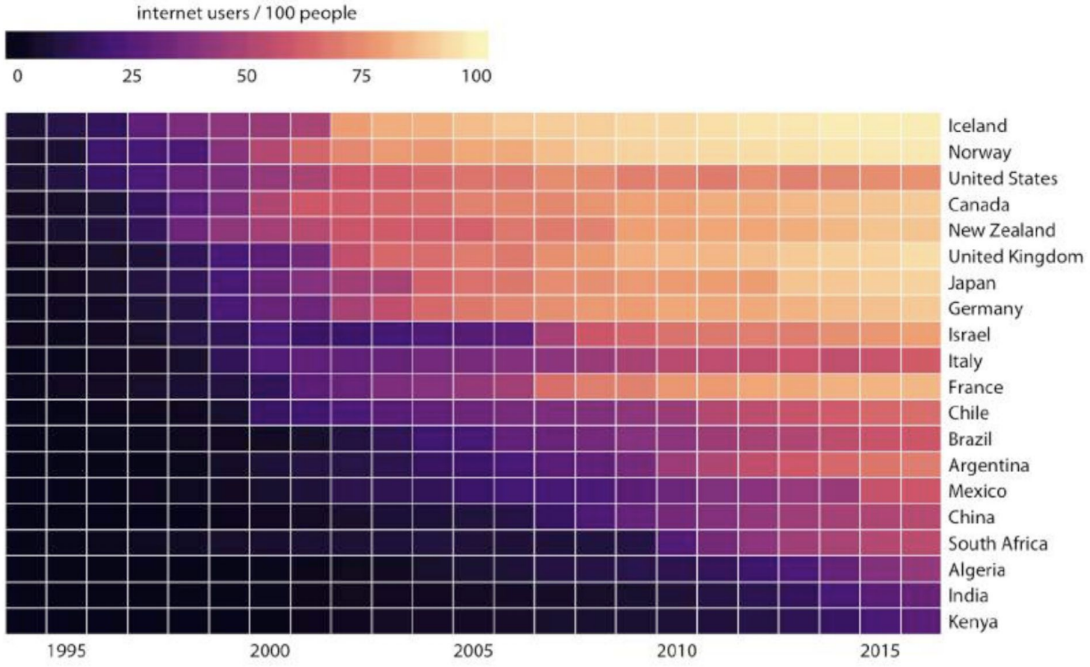


Figure 6-15. Internet adoption over time, for select countries. Countries were ordered by the year in which their internet usage first exceeded 20%. Data source: World Bank.

Reference

The notes and plots in the presentation are compiled from Claus O. Wilke's book, *Fundamentals of Data Visualization*.