Visualizing amounts

In many Scenarios, We are interested in the magnitude of some set of numbers. For examples, we might want to visualise the total sales volume of different brand of cars, or the total number of people living in different cities, or the age of olympian performing different sports. The standard visualization in this scenarios is the bar plot. In case where we have categories (eg. brands of cars, cities, or sports) and quantitative value for each category, we generally use bar plot.

The most common approaches to visualising the amounts are given below-

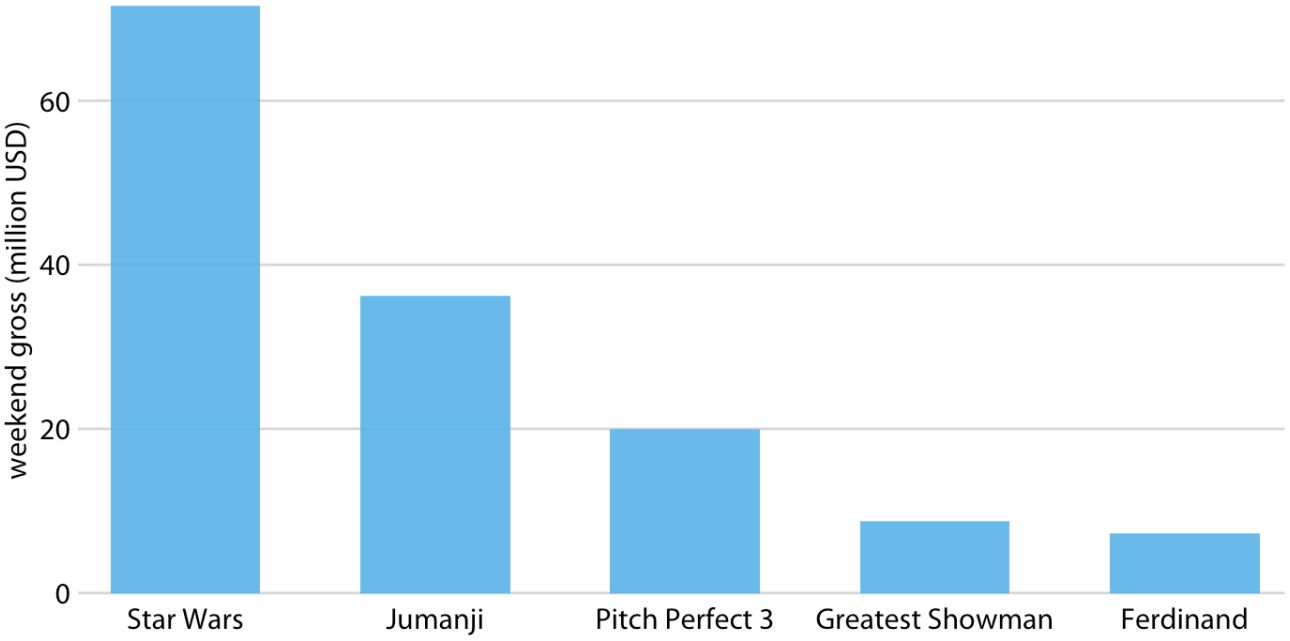
Bar plots

consider the total ticket sales for the most popular movies on a given weekend.

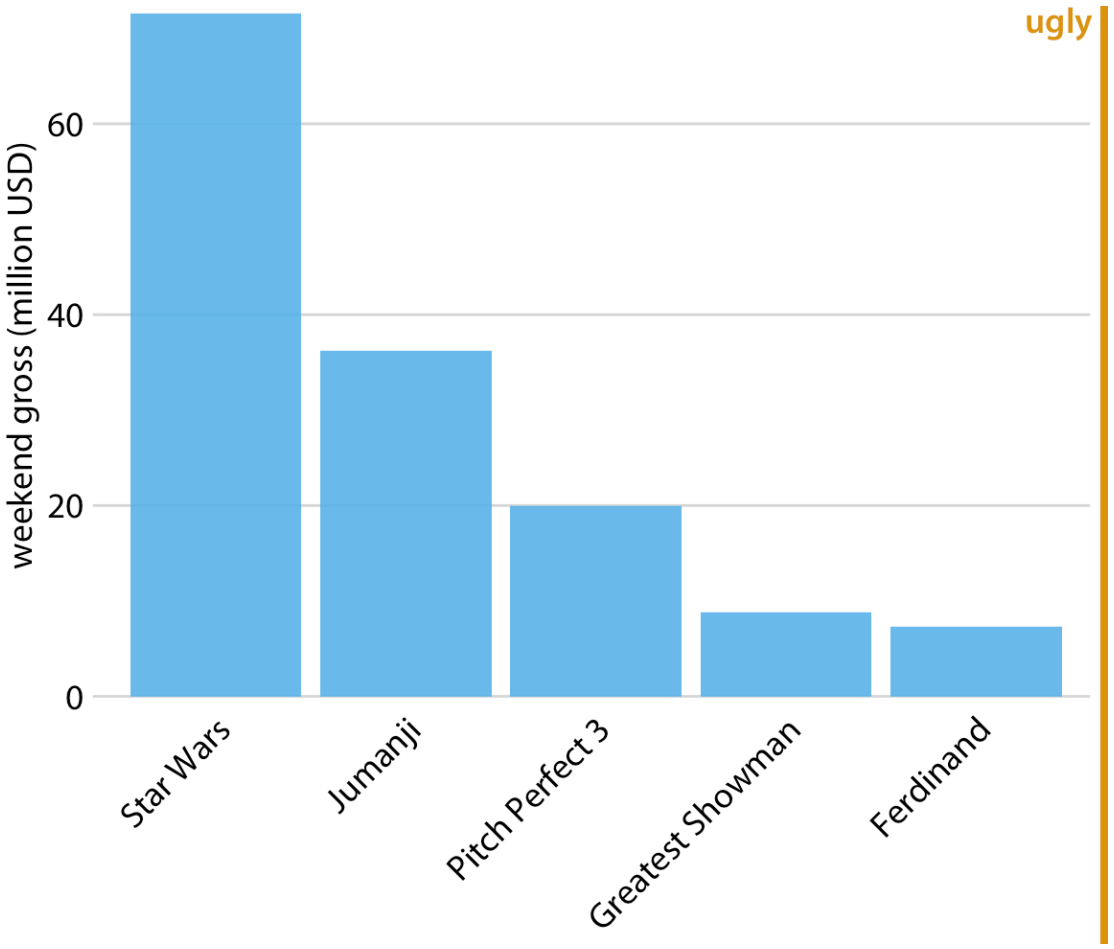
Table 6.1: Highest grossing movies for the weekend of December

Rank	Title	
1	Star Wars: The Last Jedi	
2	Jumanji: Welcome to the Jungle	
3	Pitch Perfect 3	
4	The Greatest Showman	

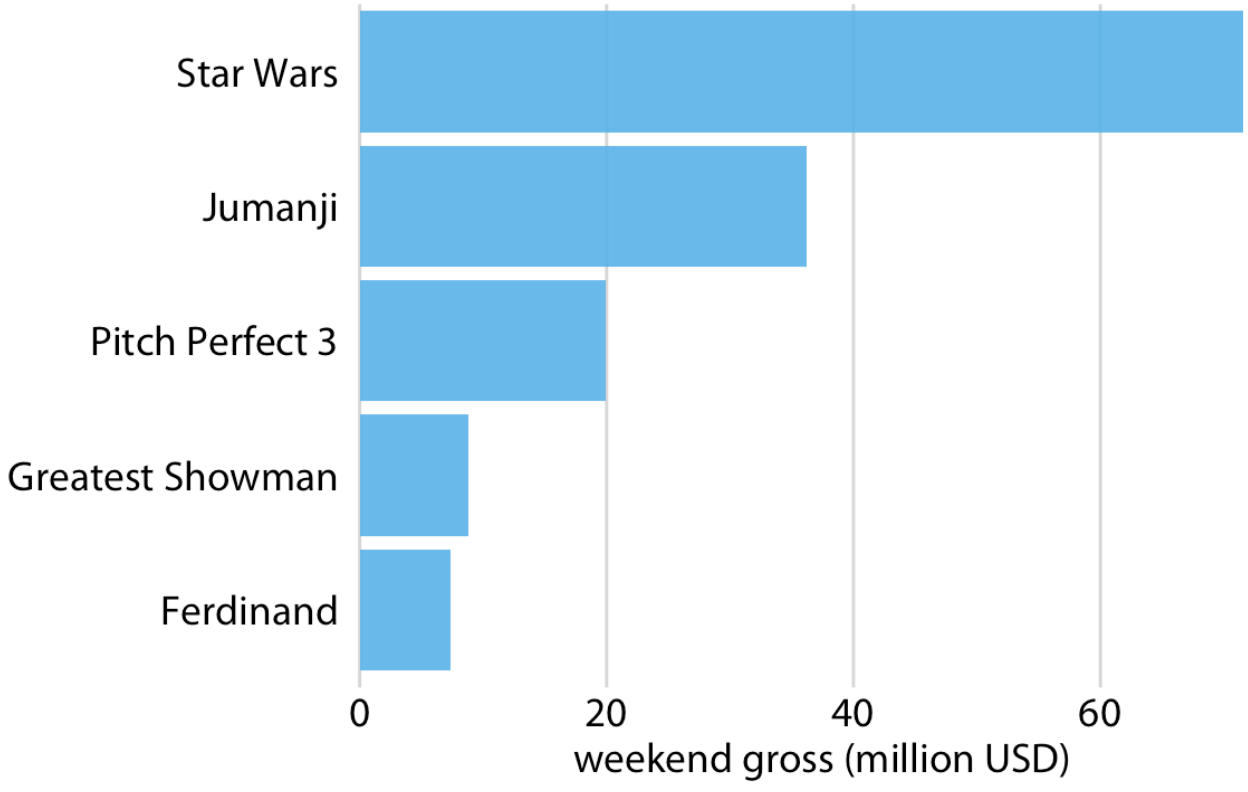
This kind of data is commonly visualised with vertical bars. For each movie, we draw a bar that starts at zero and extends all the way to the dollar for that movie's weekend gross.



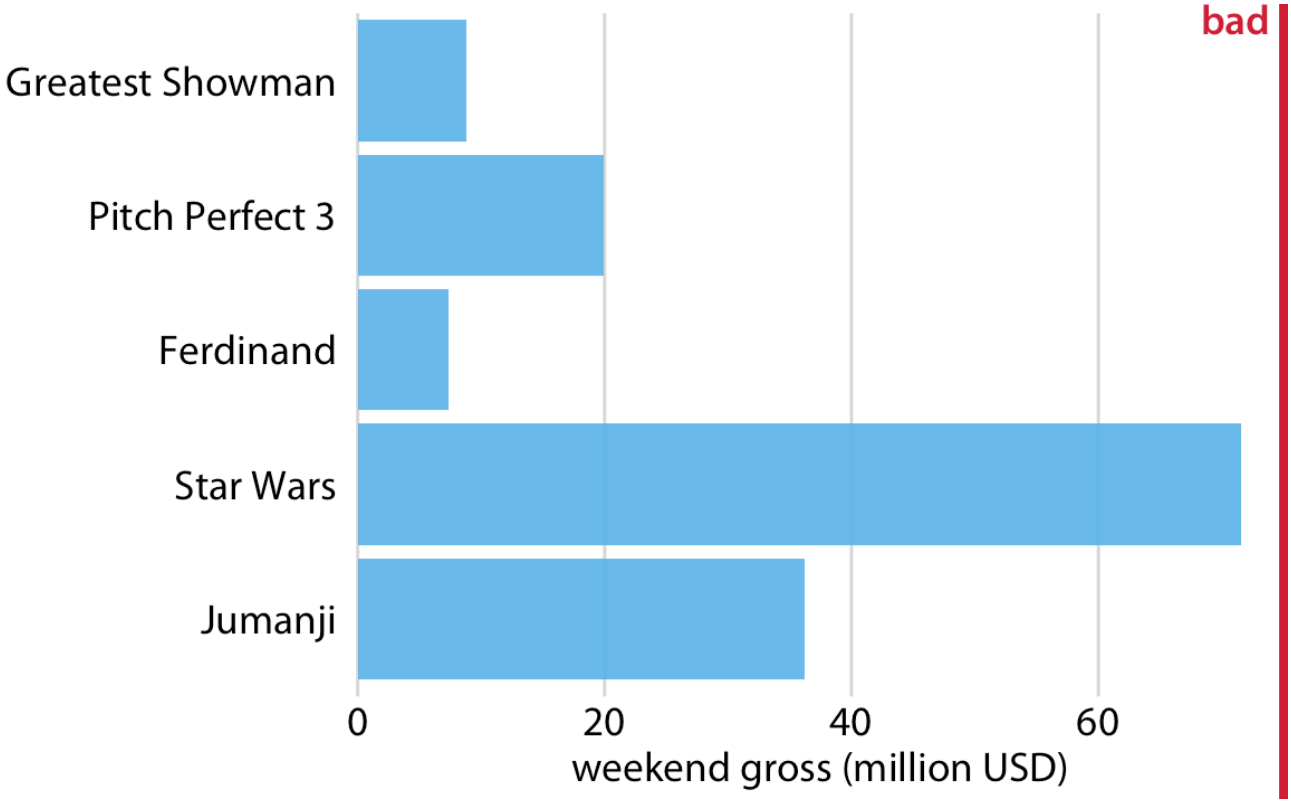
One problem we commonly encounter with vertical bars is that the labels identifying each bar take up a lot of horizontal space. To save horizontal space, we could place the bars closer together and rotate labels but this type of plot result in awkward and difficult to read. In my experience, rotated labels don't look good.



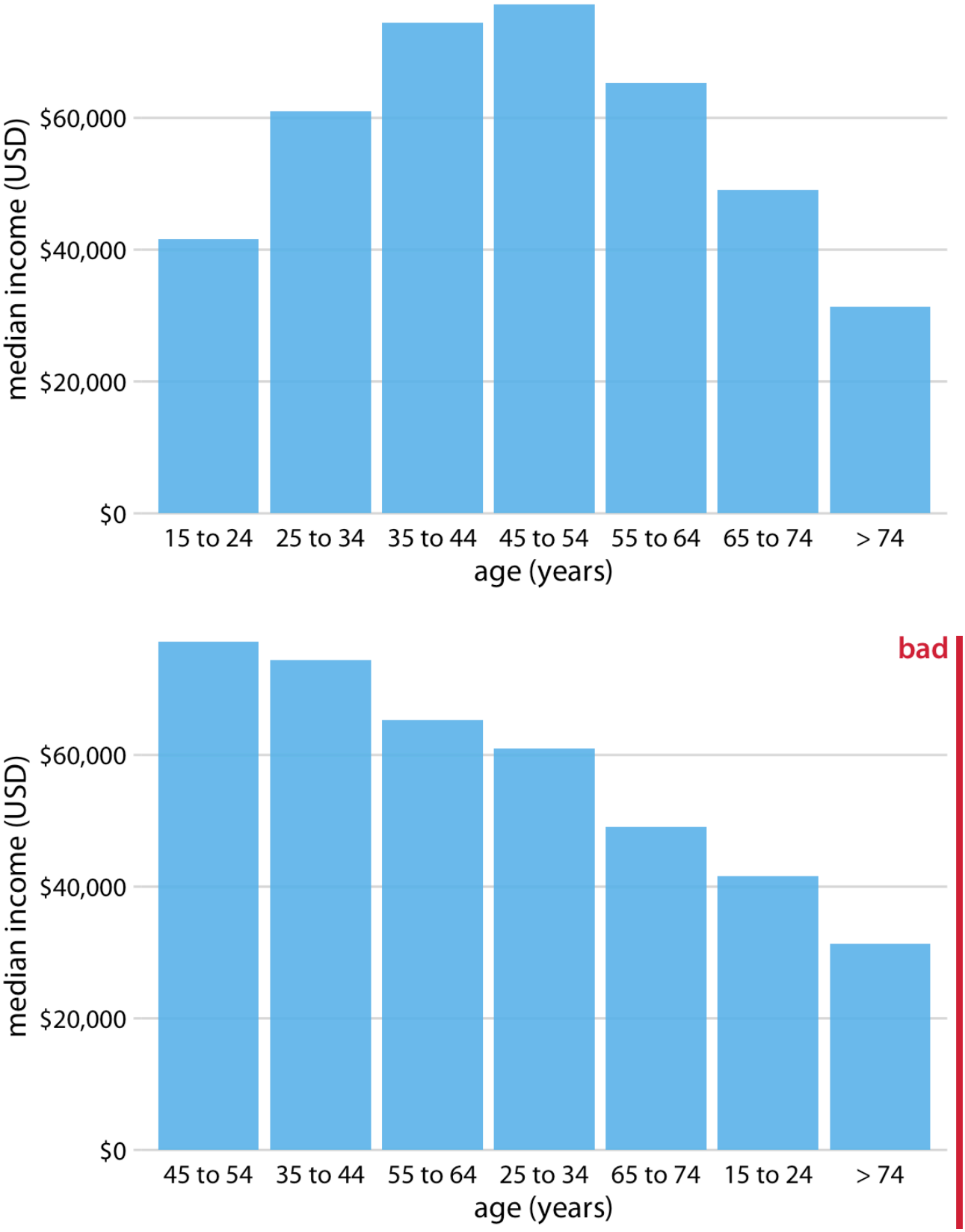
The better solution of long labels is usually to swap the x and y axis. After swiping axes we obtain a figure in which all visual elements, including all text, are horizontally oriented. As a result, the figure is much easier to read.



We need to pay attention to the order in which the bars are arranged. I often see bar plots where the bars arrange arbitrarily. Some plotting program arrange bars by default in alphabetic order of the labels and other, similarly arrangements. In general, the resulting figures are more confusing and less intuitive than figures where bars are arranged in order their size.



We should rearrange bars, when there is no natural ordering to the categories. If there is natural ordering we should retain the order in visualisation. For example the given below figure shows the annual income of different age groups. In this case, the bars bar should arrange in order of increasing age. Sorting by bar height makes no sense here.

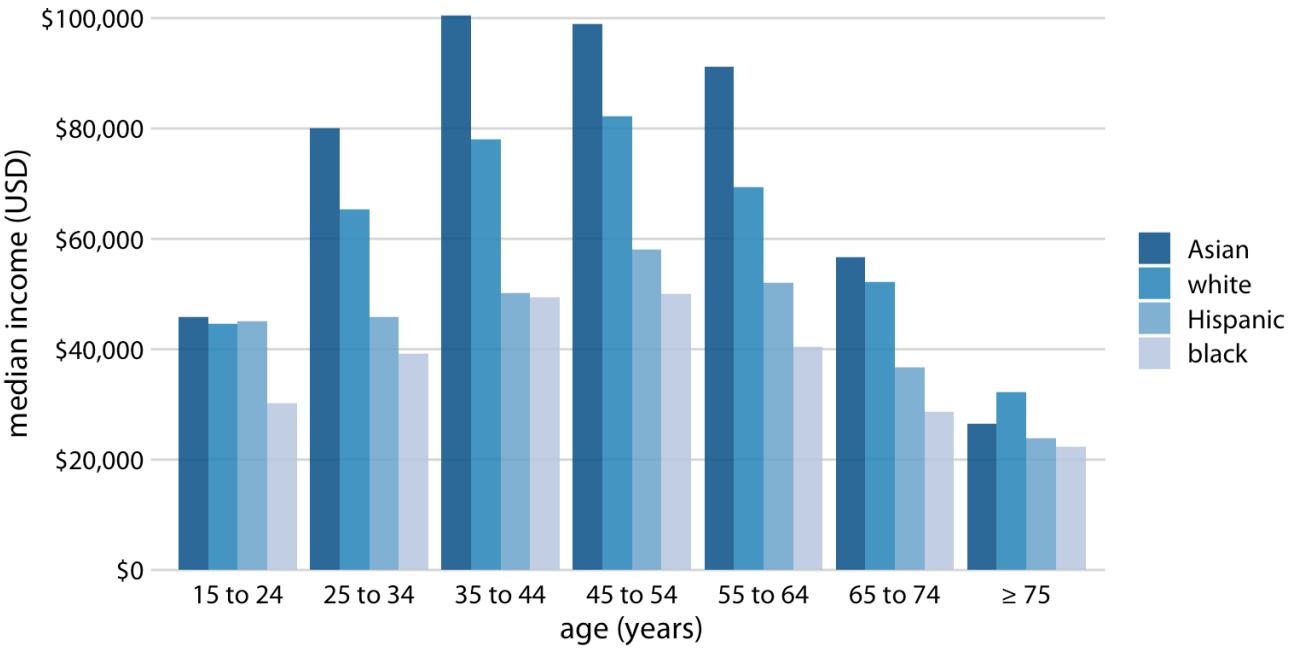


Note-

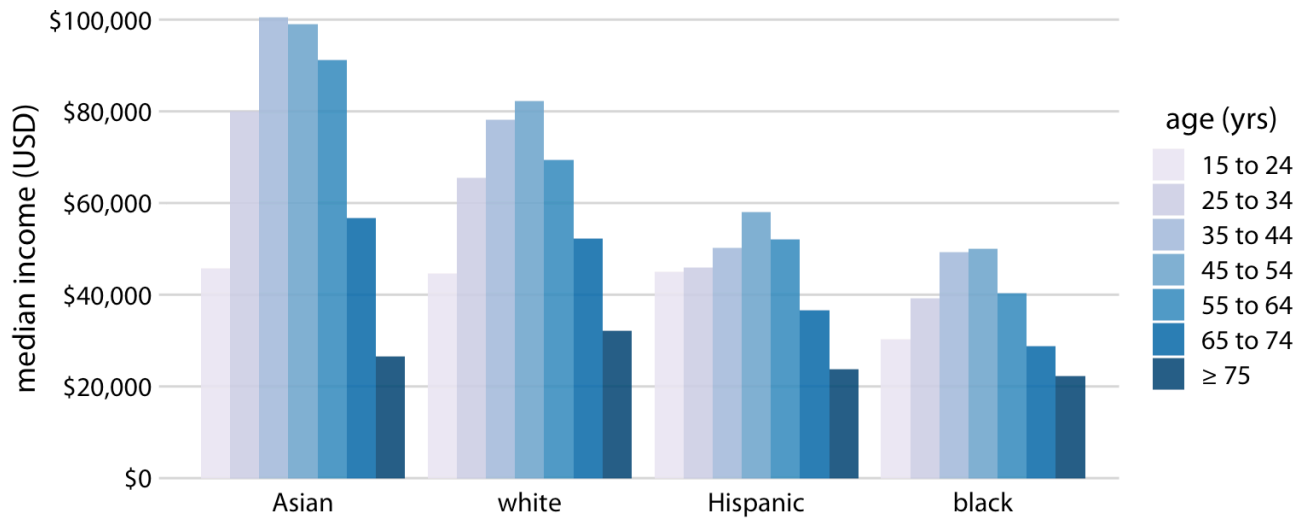
Pay attention to the bar order. If the bars represent unordered categories, order them by ascending or descending data values.

Grouped and Staked bars

However, We are interested in two categorical variables at the same time. For example, the U.S. Census Bureau provides median income level broken down by Both by age and race. We can visualise this dataset with grouped bar plot. In a grouped bar plot, we draw group of bars at each position along the x axis, determined by one categorical variable, and then we draw bars within each group according to the other categorical variable.

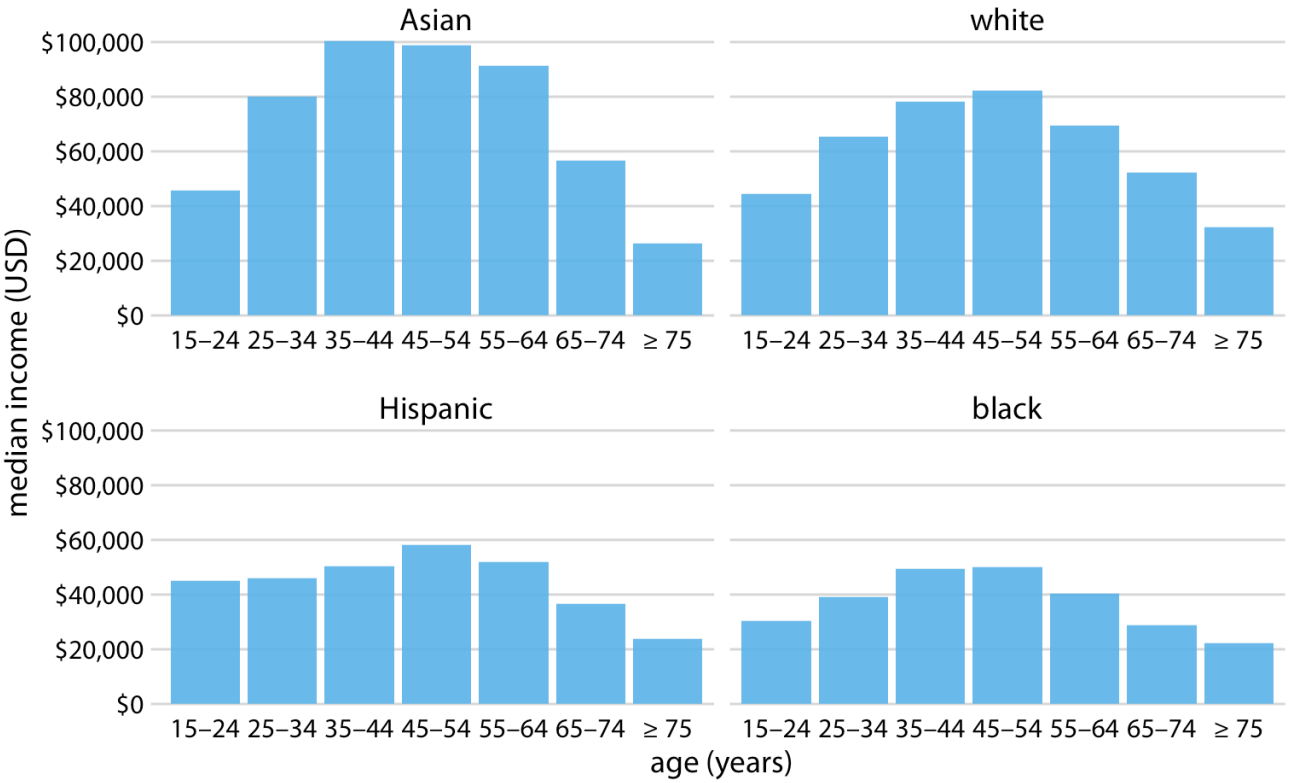


Grouped bar plots shows a lot of information at once and they can be confusing. I find the above figure difficult to read. In particular, it is difficult tot compare median income across ages groups for given racial groups. Above figure is only appropriate if we are primarily interested in the differences in income levels among racial groups, separately for specific age groups. If we care more about the overall pattern of income levels among racial groups, it may be preferable to show race along the x axis and show ages as distinct bars within racial group.



Above Figure- now race is shown along the *x* axis, and for each race we show seven bars according to the seven age groups.

In above plots we plot one categorical variable along x-axis and one we plot by colours. In my opinion plating by position is easy to read while plotting by bar colour requires more mental effort, as we have to mentally match the colours of the bars against the colours in the legend. We can avoid this mental effort by showing four separate regular bar plots rather than one grouped bar plot, ultimately this is the matter of taste.



Instead of drawing groups of bars side-by-side, it is sometimes preferable to stack bars on top of each other.
Stacking is useful when the sum of the amounts represented by the individual staked bars is in itself a meaningful amount.

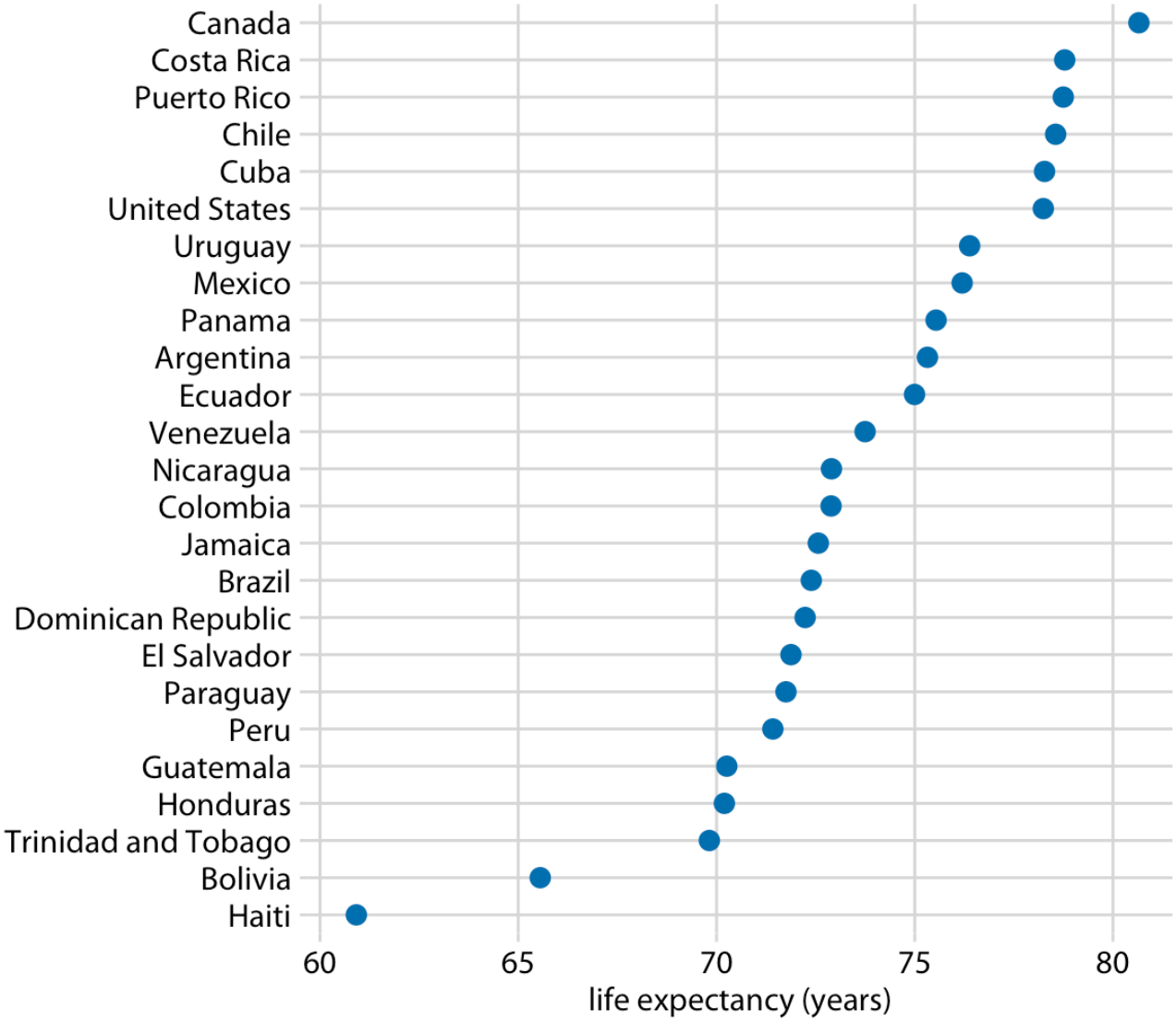
Let's take an example of passengers on Titanic, which sank on April 15, 1912. On board were approximately 1300 passengers. The passengers were traveling In one one of three classes(1st, 2nd, 3rd), and there were almost twice as many male as female passengers on the ship. To visualize passengers by class and gender, we can draw stack bar representing women on top of the bars representing men. The combined bars represents the total number of passengers.

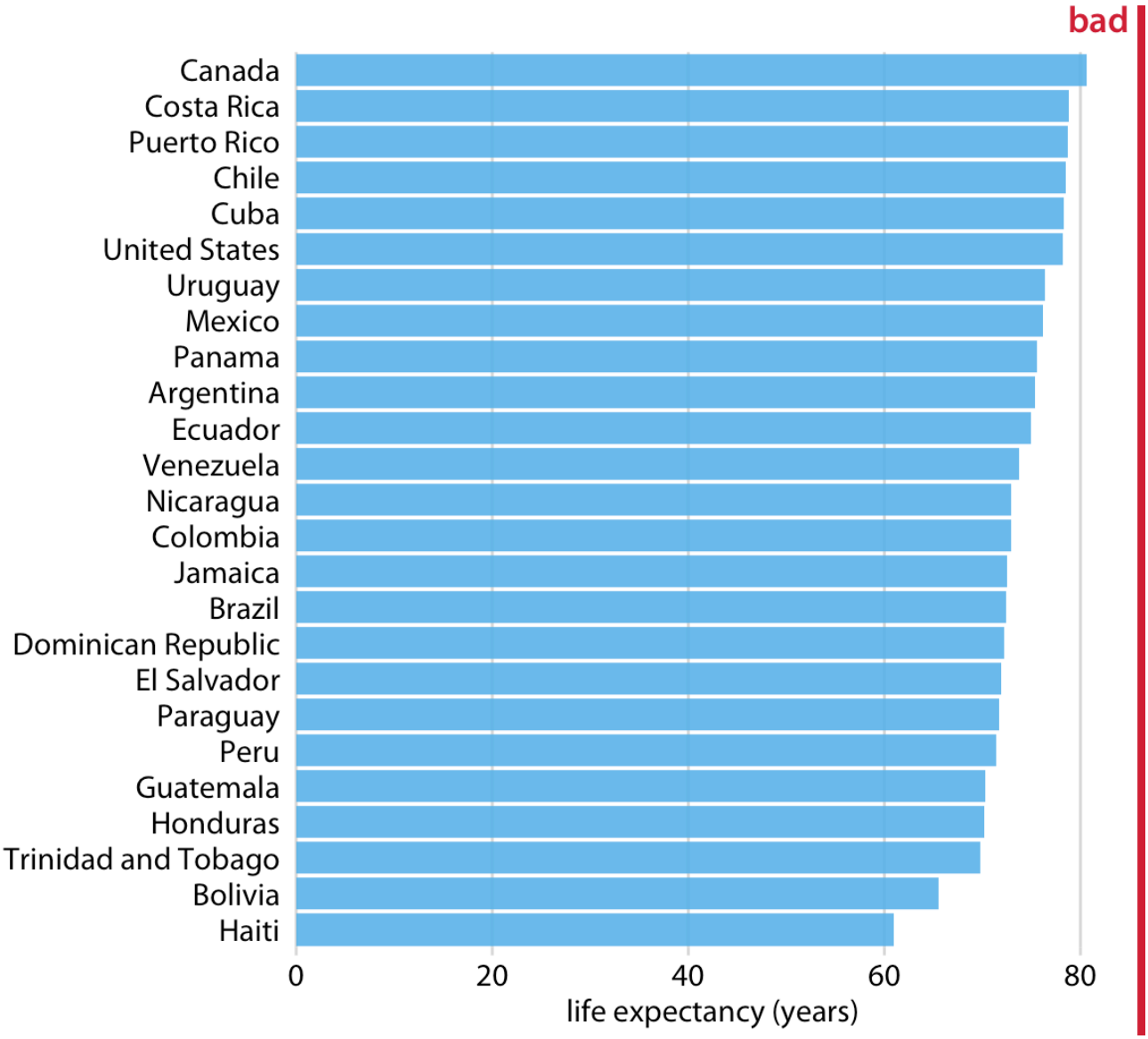
Dot plots and heat maps

One important limitation of bars is hat they have need to start at zero, so that the bar length is proportional to the amount shown. For some dataset it is Impractical

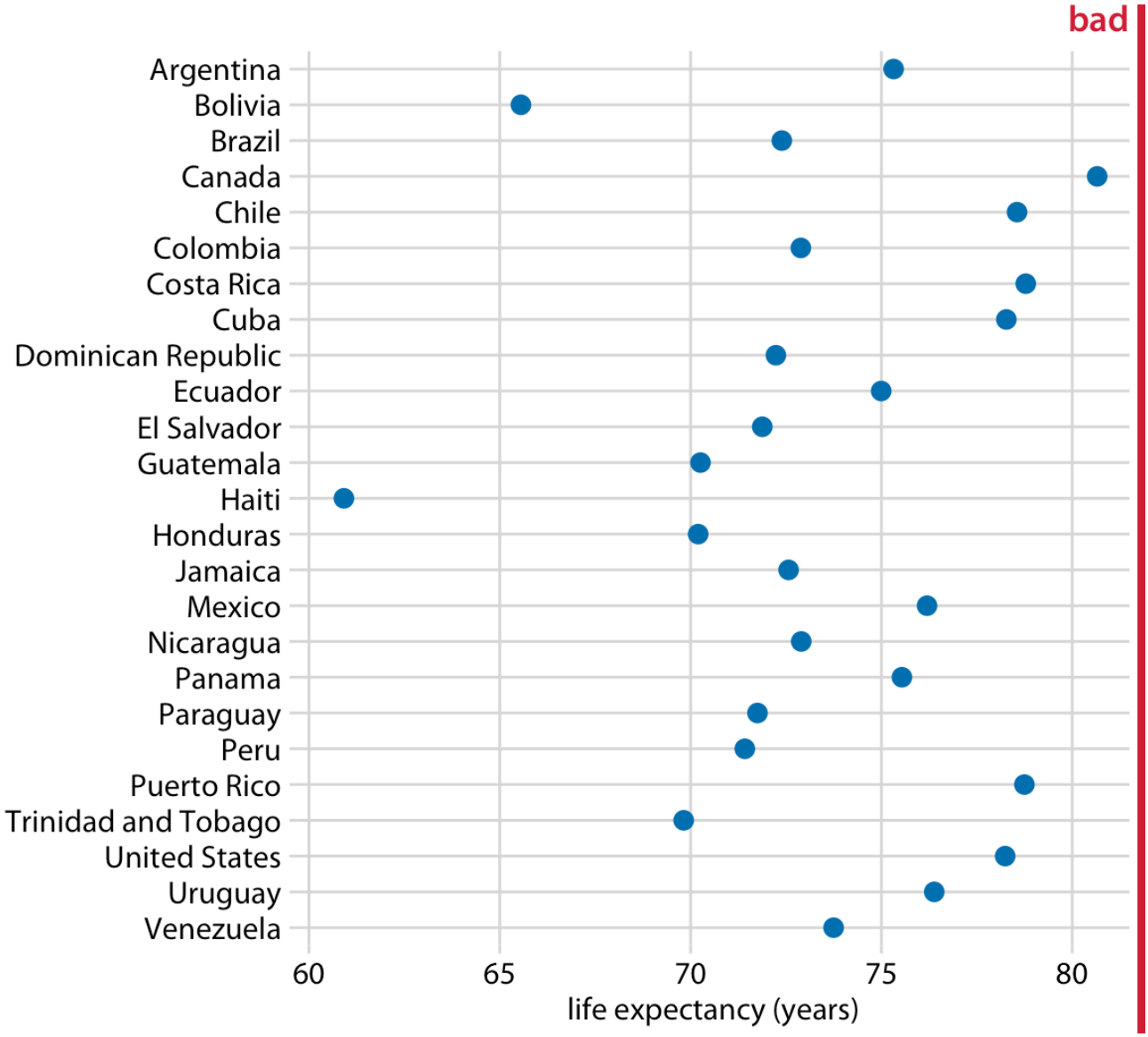
or may obscure key feature. In this case we can indicate amount by placing dots at the appropriate location.

Eg. figure given below shows life expectancy of 25 countries. The citizens have life expectancy between 60 and 81 years, for this we are limiting the x-axis range to the interval from 60 to 81 years. As we can see Canada has the highest life expectancy among all listed countries, and Bolivia and Haiti have much lower life expectancy than all other countries. If we used bars instead of dots the graph will become obscure. Because the bars are so long in this figure, and they all have nearly the same length, the eye is drawn to the middle of the bars than to their end points and the figure fails to convey its message.

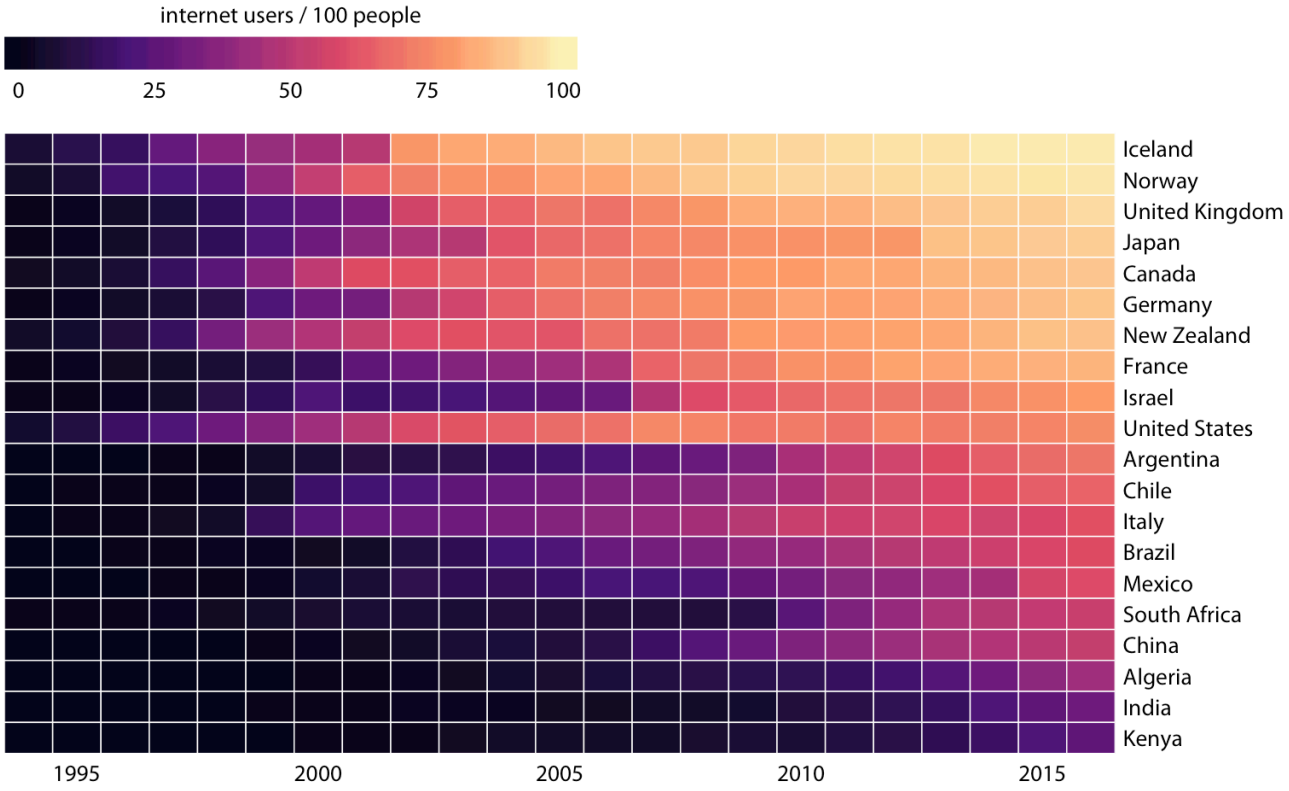




However, we need to pay attention to the order of data values the countries are are ordered In descending order of life expectancy. If we instead ordered them alphabetically, we'd end up with disordered clouds of points that is confusing and fails to convey a clear message.



As an alternative to mapping data values onto positions via bars or dots we can map data values on colours such a figure is called *heatmap*.
For example the figure below show the internet use from 1914 to 2016. In the figure we can clearly see that which country Internet use began early which did not, and we can also clearly see which countries have high internet penetration in the final year.



About Figure - Internet adoption over time, for select countries. Colour represents the percent of internet users for the respective country and year. Countries were ordered by percent internet users in 2016.

Conclusion - valid data representation is depends on the story we wants to convey by the plot.