Aalto-yliopisto
Aalto-universitetet
Aalto University

CS-E4840

Information Visualization

Assignment 1

Aitor Urruticoechea Puig
aitor.urruticoecheapuig@aalto.fi
Student N°101444219

March 2024

Contents

List of Figures

1 Exercise 1

1.1 Analyze the visualization starting from Tufte's principles

Analyze the visualization in Figure 1, starting from Tufte's principles. List at least four items that contradict these good-design principles.

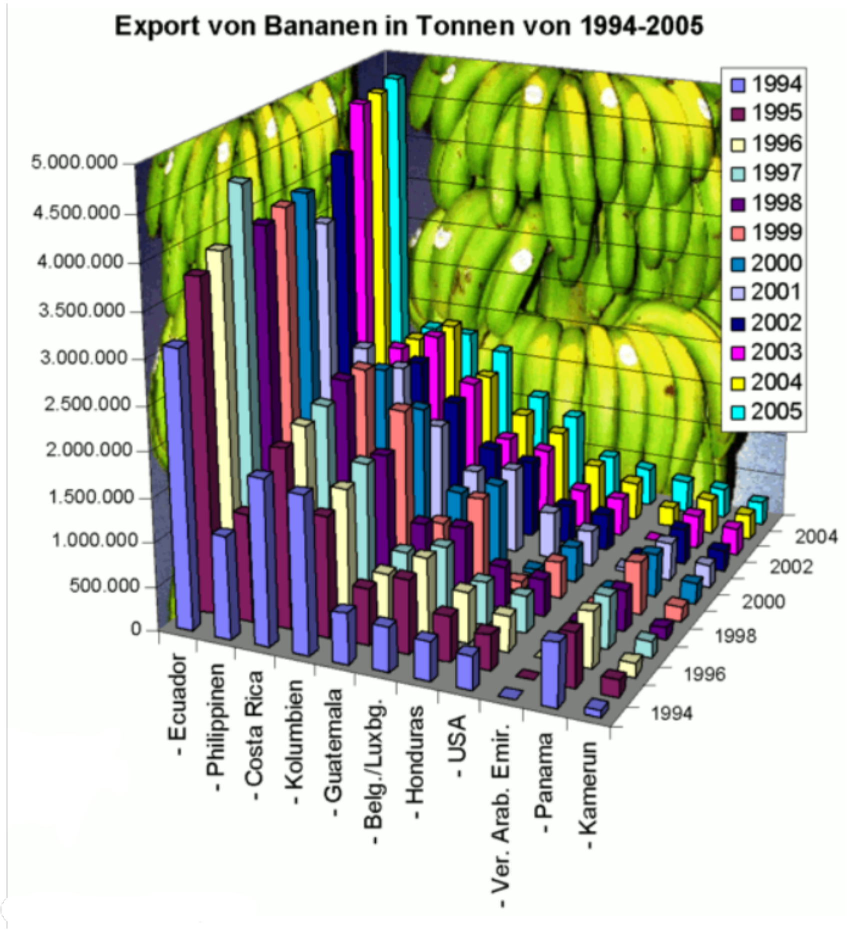

Figure 1: Statistics of banana export.

Working with Tufte's principles, one must keep in mind the main goals of following them, which is to convey the greatest number of ideas (or most information, or similar) in the shortest time, with the least ink, and in the smallest space. Intuitively, Figure 1 fails in every aspect: the graphic requires actual effort by the viewer to understand the information, has almost no ink-free space, and feels massive in a page. Let us, however, analyse in depth the main pitfalls that makes this graphic so poor from an information visualization point of view:

- **The least ink**: this graphic contains an enormous proportion of redundant data-ink. One of the main variables, the year, is represented twice. Once in the z-axis / depth axis, with its own labels; and a second time with the colour of the bars, which requires a legend of its own. Added to that, the figure contains massive amounts of non-data-ink: from the grey "floor" of the graphic to the banana background.
- **Chart junk**: the aforementioned banana background not only provides no extra information, but also doubles (and triples) down as a distraction and making data markers harder to read.
- **Poor to non-existent data integrity**: The unnecessary use of a 3D figure forces the author to

play with perspective, which can easily fool the uninformed viewer by displaying bars that are "further away" smaller. To make matters worse, 3D bars displayed in a 2D plane can easily appear hidden or partially concealed, as is the case in this graphic.

- **Data puzzle**: Due to the fact that the years are displayed redundantly, as an axis and as a colour scale, the visualization requires extra effort on the part of the viewer to understand what is being shown. Added to the confusion, while recent years are displayed in the background, resulting in them being closer to the top of the graphic; the year-to-colour legend is shown top-to-bottom, with the oldest year at the top; which makes the intuitive correlation even harder.

1.2 Suggest an improved visualization

Suggest an improved visualization and explain your design choices. For a full mark, you should provide an image (e.g., drawing, even by hand) and explain why your proposal is better than the original.

After some iteration, the improved version can be consulted in Figure 2. While the data that is shown is only a part of the one shown in Figure 1, extracting data from there was an impossible task due to the poor readability of that graphic. Matter-of-fact, the data represented is an estimation that has been extracted from the provided figure, and should not be taken at face-value. Nevertheless, it is easy to imagine how the proposed figure can be expanded to include the rest of the countries without making it unreadable.

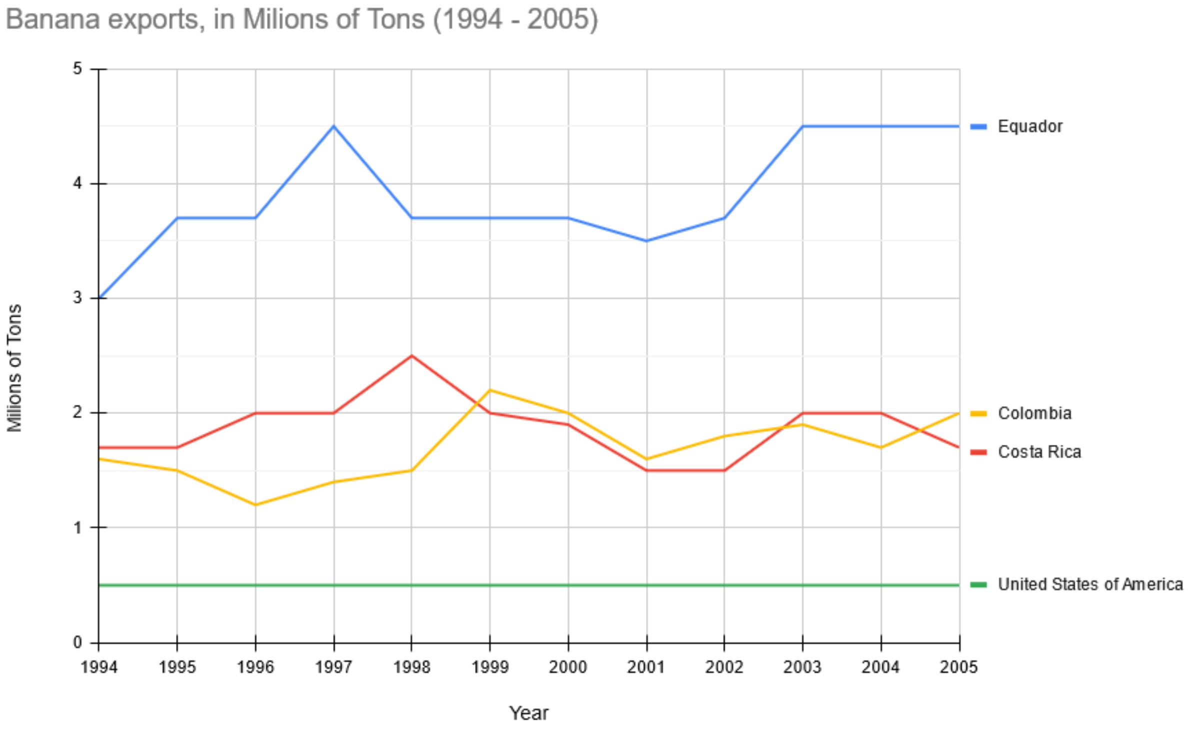

Figure 2: Statistics of banana export, improved.

The main design philosophy that has been followed to provide this improvement is the "de-cluttering" of the original figure; meaning a drastic reduction of the amount of ink. Unnecessary backgrounds, legends, and extra axis have been removed to provide a simple 2D production-vs-time graphic. Bars have, in turn, been replaced by simple and more readable lines, which provide multiple advantages. Less ink is used to represent the same data, individual data points can be more easily traced to the respective axes, and colours can be reassigned to represent the different countries without making the graphic less readable. This results in individual country progressions can be more easily followed, while cross-country comparisons can now be done with the necessary context of past and future years.

Finally, and importantly as well, "Tons" have been replaced by "millions of Tons"; which allows for a simpler notation on the vertical axis; without having to resort to the usage of many zeroes for each number. A softer-gray, secondary marker line has been provided for the half-ton increases, which is intended to help in the readability of the data points that are further away from the primary marker lines without cluttering the figure with data-ink that can potentially make the graphic harder to read.

2 Exercise 2

2.1 Help Satoshi convince the public that Bitcoin has performed better than the the Nasdaq.

Satoshi is running a crypto business. It is very turbulent with fake media is spreading rumors of bubbles and pyramid schemes. Your goal is to help Satoshi convince the public that Bitcoin has performed better than the the Nasdaq. Use the provided data (Bitcoin.csv and Nasdaq.csv), which contains the daily closing prices in US dollars for the Bitcoin and the Nasdaq-100 index, espectively, to make your case. You can every trick in your book: chartjunk, optical illusions, "creative" layout, use only part of the data. You can use any plotting software available (R, Matlab, Python, Excel, OpenOffice, gnuplot etc.)

The main goal of Figure 3 is to showcase Bitcoin as the ever-growing winning phenomenon against a stagnant NASDAQ index, unable to even produce minimum growth. Using some chartjunk, clever data cutting in both axes, and an exaggerated visualization that favours Bitcoin by using bars instead of a line make it look like it is basically obvious that a Bitcoin investment is the smart move.

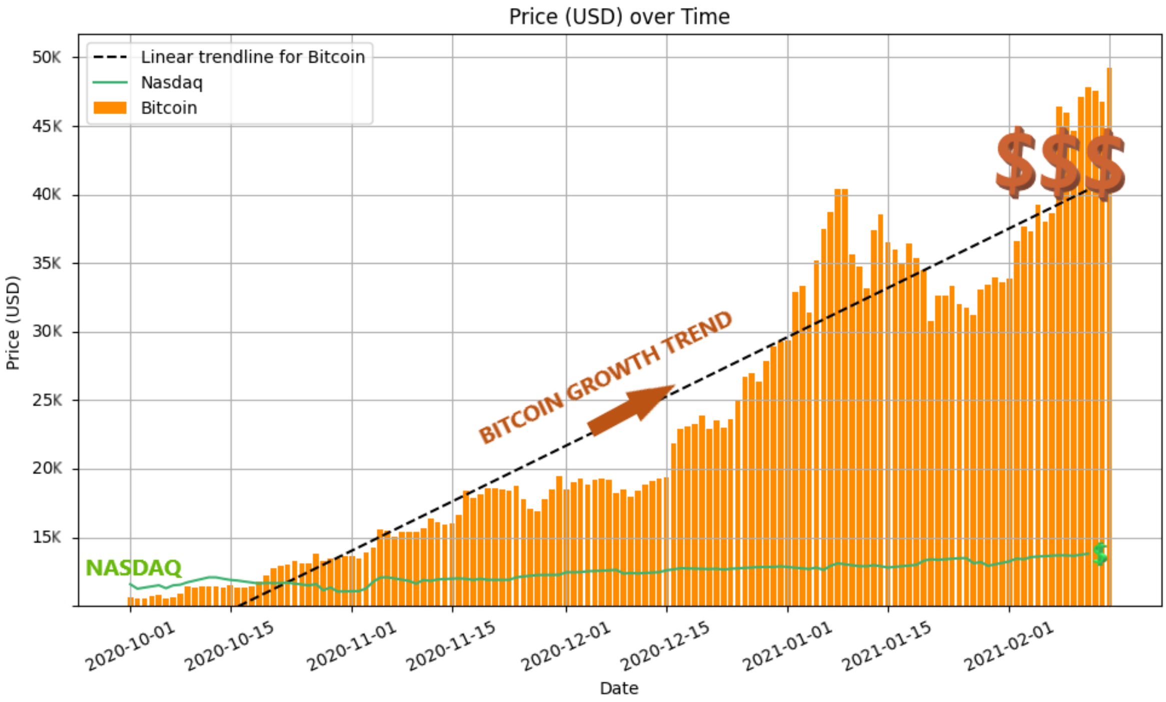

Figure 3: Bitcoin vs. NASDAQ stock prices, with visualization bias in favour of the former.

2.2 Help Warren convince the public of the contrary

Warren is a passive investor irritated by the whole Bitcoin fuzz. Use the same data to make the opposite case. Again, you can use every creative trick imaginable.

The goal of Figure 4 is to make Bitcoin look childish and unpredictable; while traditional stock values like NASDAQ are a mature and stable alternative for more serious investors. Chart junk is included, but to a lesser extent, and the time window that has been the focus is wider to showcase stability and long-term thinking. The big trick beyond the clever data cutting is the usage of two separate vertical axes for the same figure. By making both lines start at the same point, the Bitcoin fall can be dramatized even further, while exaggerating the perceived stability and steady growth of NASDAQ; while always using different scales for each!

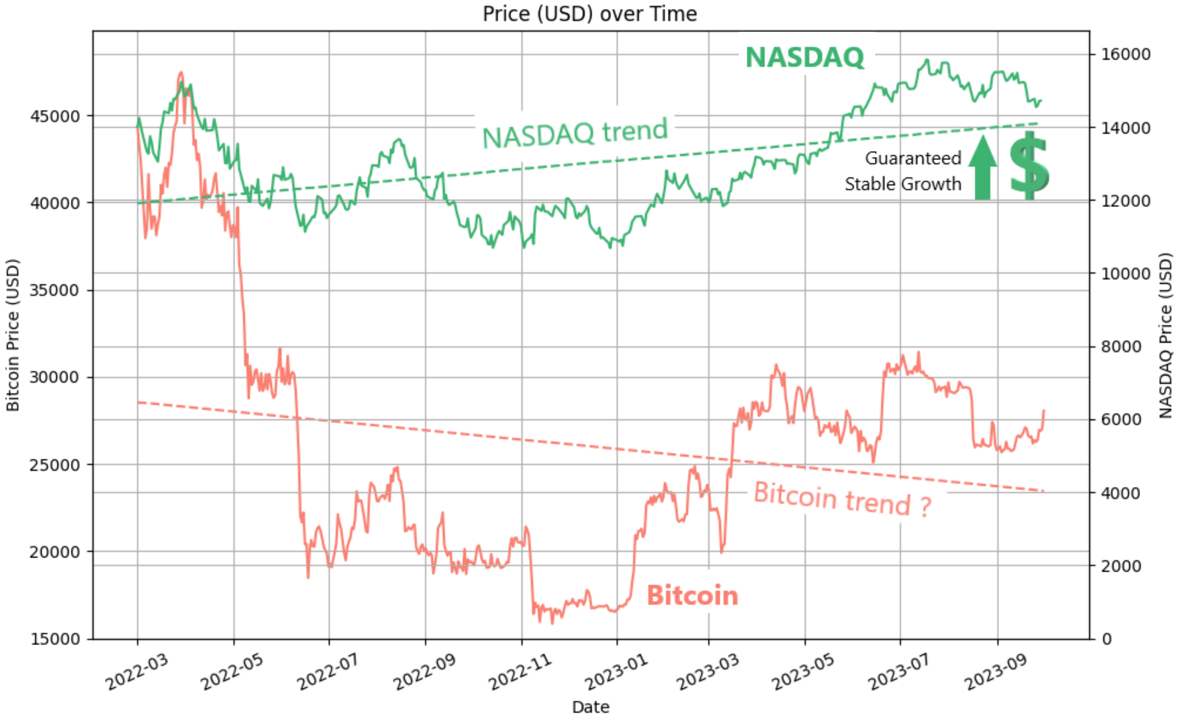

Figure 4: Bitcoin vs. NASDAQ stock prices, with visualization bias in favour of the latter.

2.3 Lie factor

Use the notion of Lie factor (see slides of Lecture 2 or Tufte's book, page 57–58) to measure whether the above plots are underestimating or overestimating the relative performance of the two financial instruments.

The lie factor can be calculated as:

$$\text{lie factor} = \frac{\text{size of the effect shown in the graphic}}{\text{actual effect in the data}} \tag{1}$$

Since the general trick used in both cases is using time windows that favour different views, the lie factor will only be to account for part of the illusion. Still, it will provide us with a numerical value that can be used to get a rough idea of how misleading the provided figures are. Focusing on Figure 3:

$$\begin{aligned}\text{lie factor}_{\text{Bitcoin}} &= \frac{\text{shown Bitcoin growth}}{\text{real Bitcoin growth}} \\ &= \frac{3100\%}{900\%} = 3.44\end{aligned} \tag{2}$$

$$\begin{aligned}\text{lie factor}_{\text{NASDAQ}} &= \frac{\text{shown NASDAQ growth}}{\text{real NASDAQ growth}} \\ &= \frac{133\%}{110\%} = 1.21\end{aligned} \tag{3}$$

While for Figure 4:

$$\begin{aligned}\text{lie factor}_{\text{Bitcoin}} &= \frac{\text{shown Bitcoin growth}}{\text{real Bitcoin growth}} \\ &= \frac{-58\%}{-39\%} = 1.49\end{aligned} \tag{4}$$

$$\begin{aligned}\text{lie factor}_{\text{NASDAQ}} &= \frac{\text{shown NASDAQ growth}}{\text{real NASDAQ growth}} \\ &= \frac{20\%}{7\%} = 2.86\end{aligned} \tag{5}$$

2.4 Help Jorma and follow the principles of Tufte as closely as possible

Jorma is a student at Aalto University. He is impartial because he has no cash, Bitcoins, or Nasdaq ETFs. He decides to start a blog about graphic design and data visualization. Help Jorma and follow the principles of Tufte as closely as possible, and create a plot for the relative performances of Bitcoin and the Nasdaq-100 index. Justify your choices, and describe how/whether you can improve your visualization further.

Figure 5 displays a more accurate representation of the provided data. No time windows are used, the full extent of the data is shown. The colours for the lines corresponding to each stock are now away from golden yellows, "good" greens and "bad" reds. Instead, purple has been chosen for Bitcoin, representing its novelty and recent relevance; while slate gray has been used for the classic and steady NASDAQ. Rather than dubious linear regressions, a 60-day rolling average accompanies each data set; which allows for more nuanced analysis of the general trends in each epoch. No exaggeration has been used; only one vertical axis is employed, and it does begin at zero this time, correctly grounding the data.

Further improvements could be made, however. The colours used could be seen by some as impartial; while others might prefer the data without the rolling average, or accompanied by other statistical measures. Further date points could be added to the x-axis to help in situating certain peaks and valleys, though adding more grid lines can easily result in a cluttered figure.

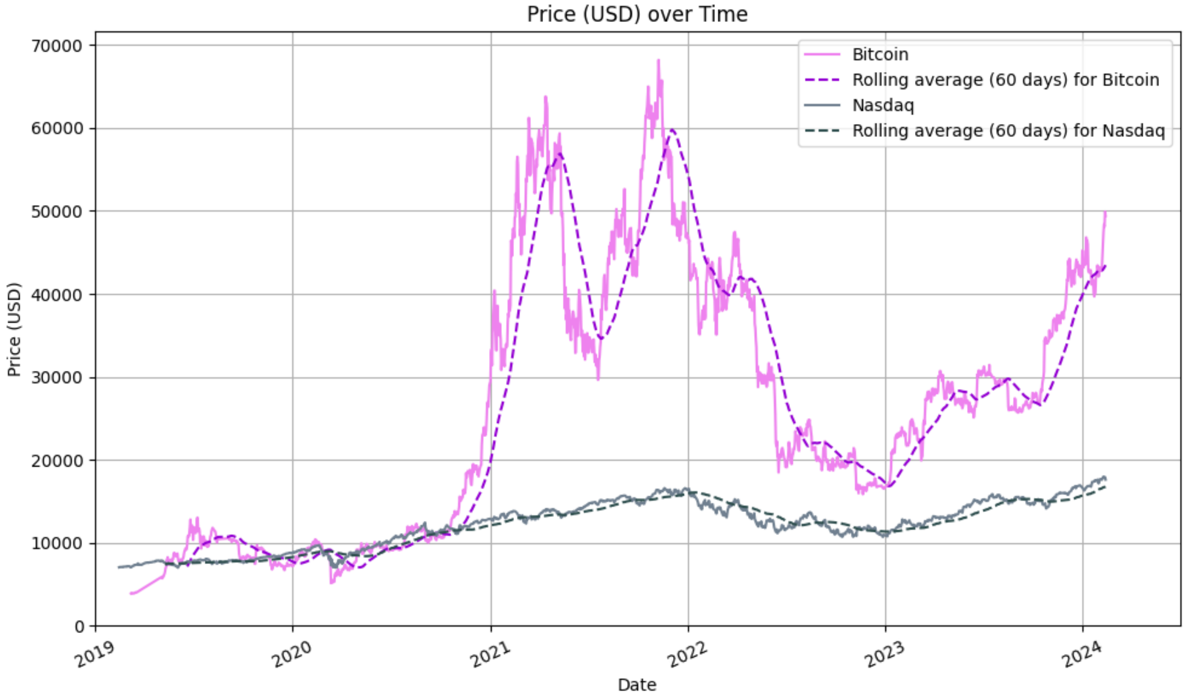

Figure 5: Bitcoin vs NASDAQ performance over time, with no bias.

3 Exercise 3

Look for an example of a visualization that you find particularly beautiful or disturbingly bad in a recent issue (published on or after June 2021) of a high-profile scientific journal (Nature, Science, etc.) or mainstream media (CNN / Helsingin Sanomat / Tilastokeskus.). Try to explain what makes it appealing, purposeful, horrible, etc. The journals are accessible from within Aalto. Insert the picture (a screenshot or photo) in the report.

Figure 6 shows, according to Spanish newspaper ABC, a figure published by the Spanish Government and Spanish national television broadcaster RTVE. According to ABC; this was published as an official graphic showcasing the results of the 2023 General Elections held in July. While initially a fairly simple graphic with easily recognisable colours and correct data with the number of seats and votes obtained for each major party; the graphic cleverly plays with the perspective and where the actual bars start to give the illusion of a different result. The bars representing the parties that formed the incumbent and now current government, the socialists "PSOE" (in red) and moderate left "Sumar" (in pink) are showcased in a way that it gives an illusion of them having more seats in congress than the conservative right "PP" (in blue) and the far-right "Vox" (in green), respectively.

The author finds this graphic particularly concerning because of the heavy implications that can be derived from the fraudulent visualization of data. The results of information visualization do not exist in a vacuum; and the context in which this are displayed and reported is crucial for understanding the whole of the information. If it is true that the government and the national television broadcaster showcased misleading information this is not a small deal. Democratic processes are at stake in many places of the world; and conveying information impartially is key in informing the voter. Yet; the "if" used previously is also key. The journal ABC is the only source where this figure has been found; and it seems like no one else is calling out neither RTVE nor the Spanish government for this attempt at conveying misleading information. This could be an oversight by other media outlets;

or all-together a ploy by ABC, famously known as a right-wing and conservative media outlet, to discredit the national broadcasting corporation and the current centre-left Spanish government.

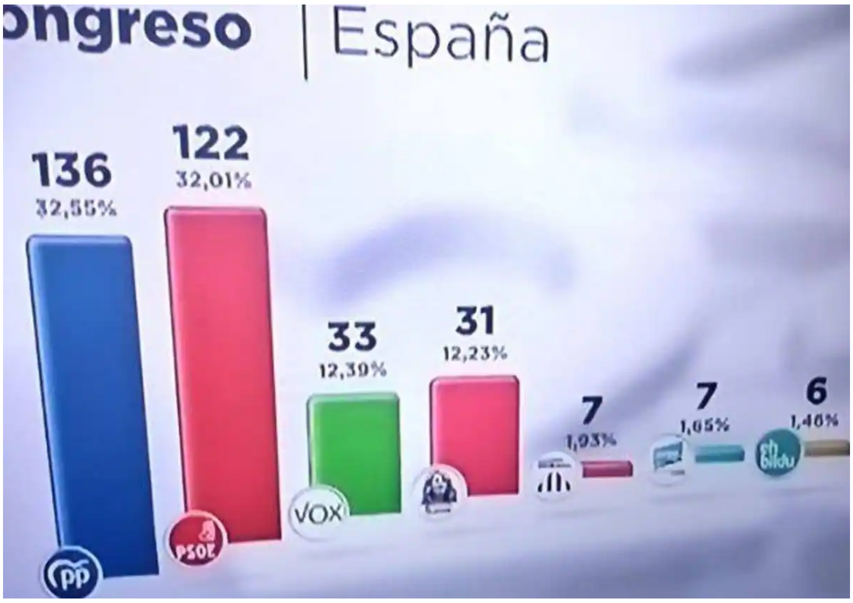

Figure 6: Spanish 2023 General Election results. Extracted from https://www.abc.es/espana/psoe-sumar-encima-pp-vox-enganosos-graficos-20230724131356-nt.html#.

All in all, the problems with Figure 6 are multifaceted to say the least. Yes, the graphic plays clever tricks to make it look like the repetition of the PSOE-Sumar coalition should be the natural result of the general election. Nonetheless, the sources for the graphic are lacking, and there is a not-to-be-discarded possibility that this manipulation is all together an attempt by a right-leaning media outlet to de-legitimate the government.

4 Exercise 4

Visualize the penguin dataset (Penguins.csv), available at the MyCourses page. This dataset contains beak measurements of 344 penguins sampled in Antarctica. Create a small-multiples (trellis) visualization with scatterplots of each pair of features, arranged as a matrix; see an example of such arrangement for the Iris dataset (see Wikipedia's "Iris Flower data set"). Indicate with different colors the three species (Chinstrap, Gentoo, Adelie). Try to show the difference between the regions, and maximize the data-ink ratio, within reason. More info about this data set (like what those measurements actually refer to) can be found here: https://allisonhorst.github.io/almerpenguins/

The proposed visualization can bee seen in Figure 7. Note that to maximize the information, the upper and lower sides of the diagonal show different highlights of the data (sex and species respectively). The islands where the data was recorded are represented by point size, while the diagonal is reserved for the histogram of each numerical value.

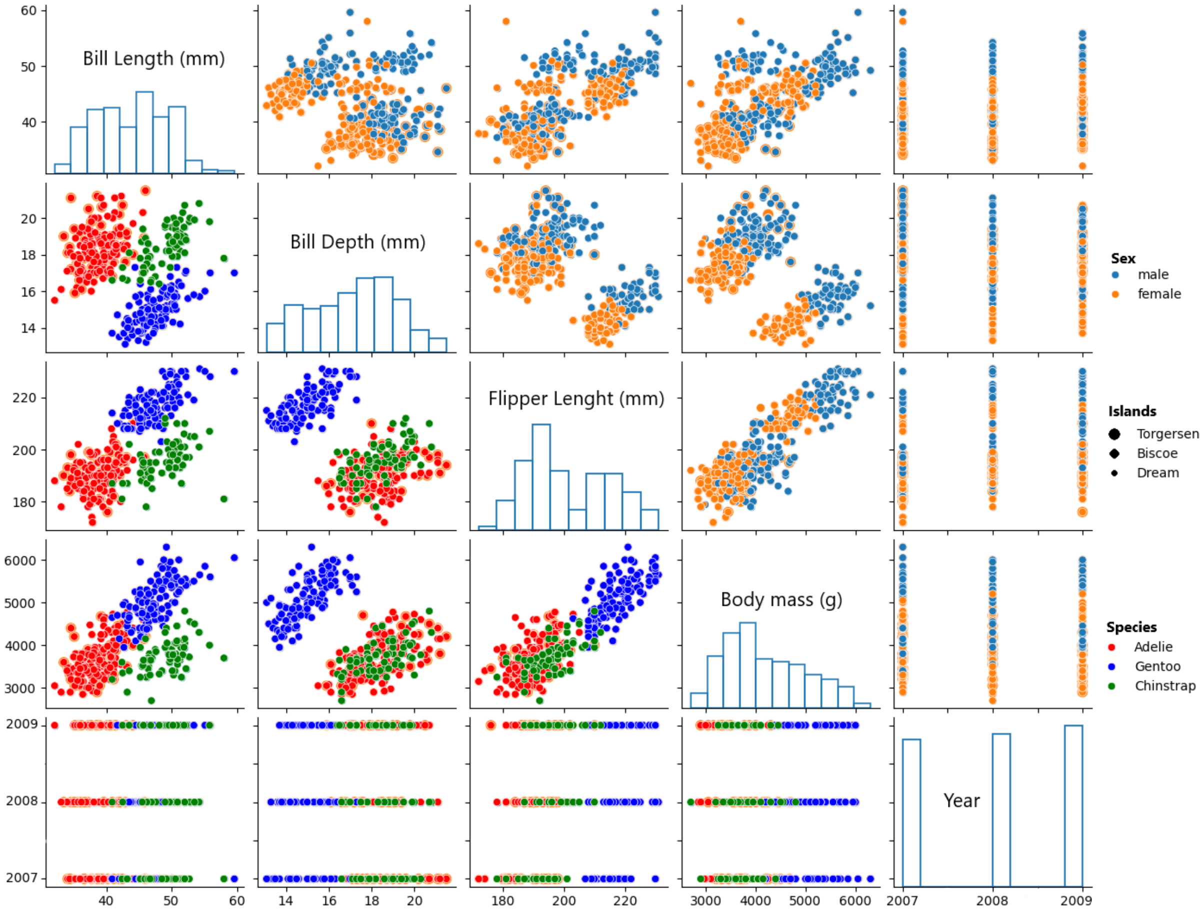

Figure 7: Trellis visualization of the given penguin population.