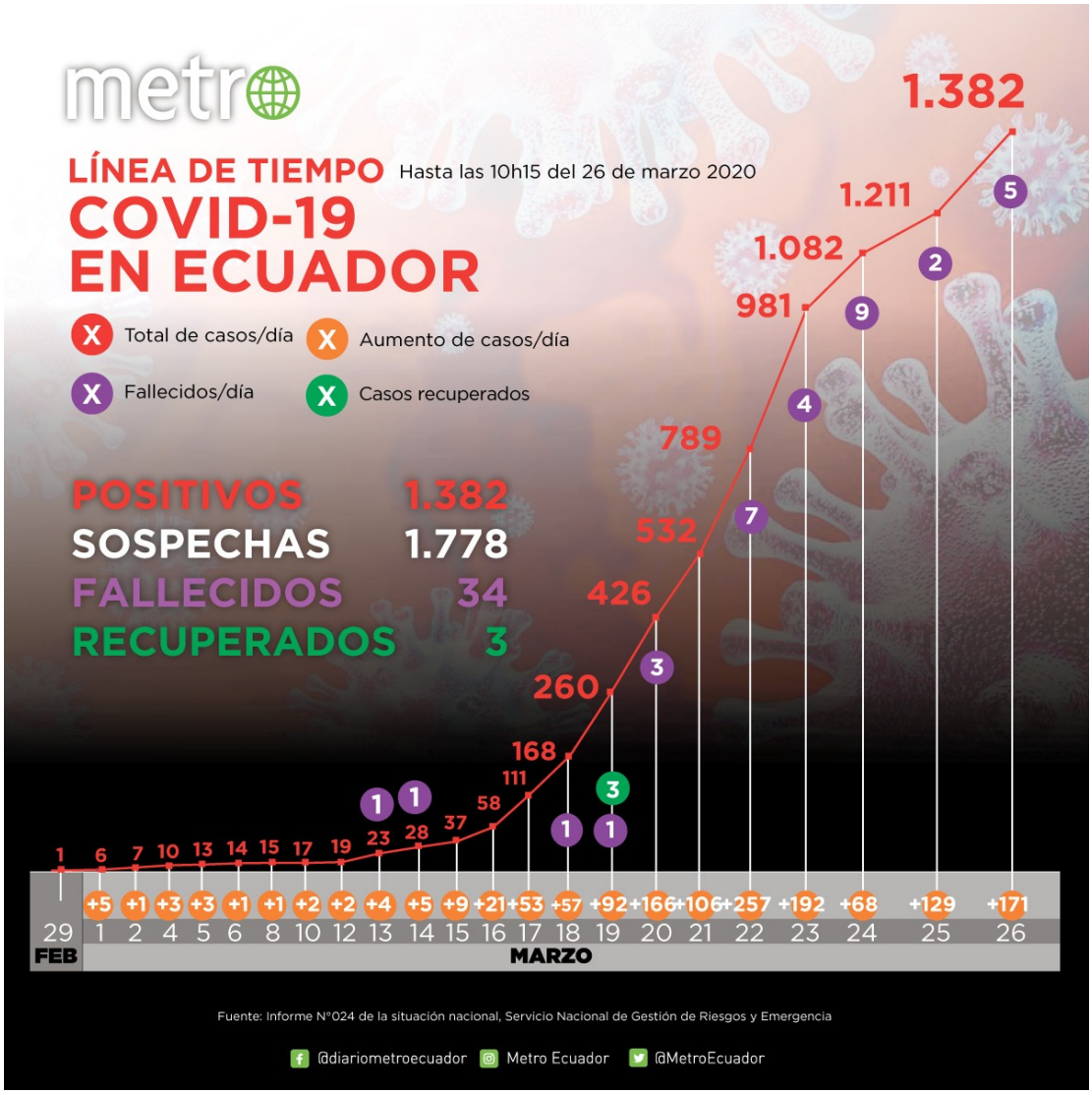
Student Name : Nguyen Xuan Binh
Student Number: 887799

Information Visualization Assignment 1

Exercise 1 (3 points)

Figures 1 and 2 show published visualizations of different issues. Your task is to:
(a) Analyze the visualization in Figure 1, starting from Tufte's principles. List at least four items that contradict these good-design principles.



Tufte's principles state that five aspects improve the visualization's quality: maximizing the data-ink ratio, reducing the chartjunk,multifunctioning graphical elements, adequate data density and using small multiples, ensuring the aesthetics and techniques. However, from Figure 1, it is noticeable that many violated Tufte's principles are violated in the graph.

Firstly, this figure violates the data-ink principle. Data-ink is the non-erasable and non-redundant core of graphics, and erasing the data-ink would reduce the amount of information transmitted by the graphics (lecture). The two main graph components in this figure are the table listing the total number of recovered and death cases, etc., and the temporal line graph that shows the number of infected cases each day in March, starting from 29th February. This table can be directly derived from the line graph, so erasing it will not reduce any information. The table is the redundant data-ink, and by erasing it, we can increase the data-ink ratio for this figure. Additionally, the vertical lines connecting the dots to the baseline are unnecessary because the total case number has already been annotated next to each data point.

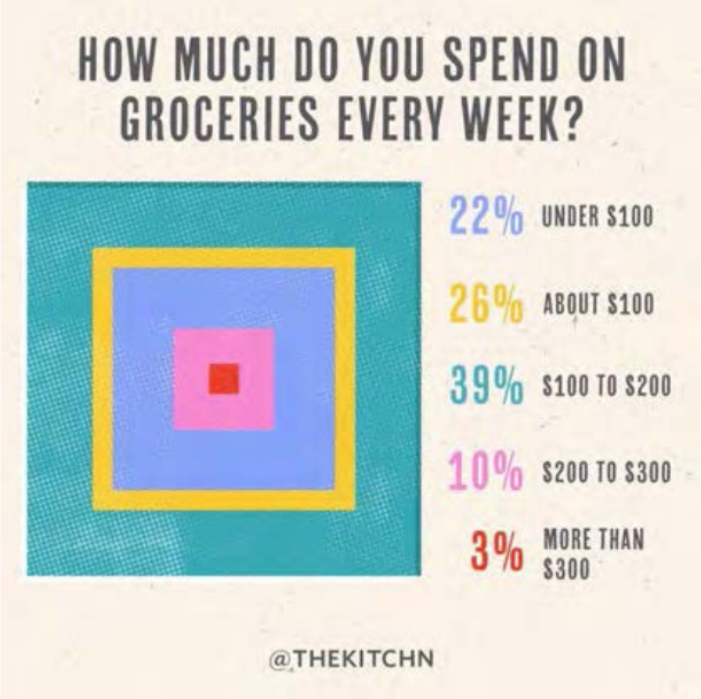
The second violated Tufte's principle is too much chartjunk. Chartjunk is the interior decoration of graphics that does not tell the viewer anything new (lecture). This figure has many redundant graphics, such as the unnecessarily large black background at the bottom, the newspaper logo Metro, the social networks icon, and the coronavirus background. These features contain no new information and distract viewers from reading the line graph. In fact, the chartjunk is so prominent that it dwarfs the actual line graph underneath.

The third principle that the visualization violates is data density and small multiples. This principle ensures that all visual distinctions are as subtle as possible but still clear and effective because large distinctions generate clutter while smaller distinctions highlight the data. However, looking at the temporal day axis, the ticks are tiny at the beginning of the month, while they become significantly further away at the end of March. This will cause an effect that from 29th February to 18th March, the data is not visually noticeable due to the larger ticks at the end of the month. When I first looked at this figure (from a viewer standpoint), I could hardly notice the data before 18th March. The last day has 1381 cases; before 18th March, the cases range from 1 - 111. This discrepancy is too large, which makes the comparison visually challenging. A better case is simply using the log scale, or using two different graphs side by side, one from 29th February to 18th March, and another from 19th March to 26th March

Finally, the figure violates the aesthetic and technique principle. The weights of the letters should be in proportion to the other visual elements, but in this figure, there are many different font sizes, some of which are very large, while some are so small that they are barely visible. Additionally, graphics should usually have a greater length than height, especially when it is a time series graph. Ideally, the golden ratio 1:1618 works best for many cases. However, this figure shape is a rectangle, where the height is longer than the width. This will cause discomfort to the viewers who are used to reading the time series graph wider in the horizontal axis.

(b) Suggest an improved visualization for Figure 2, using the data shown in the figure, 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.

Figure 2. Grocery expenditure according to a survey.

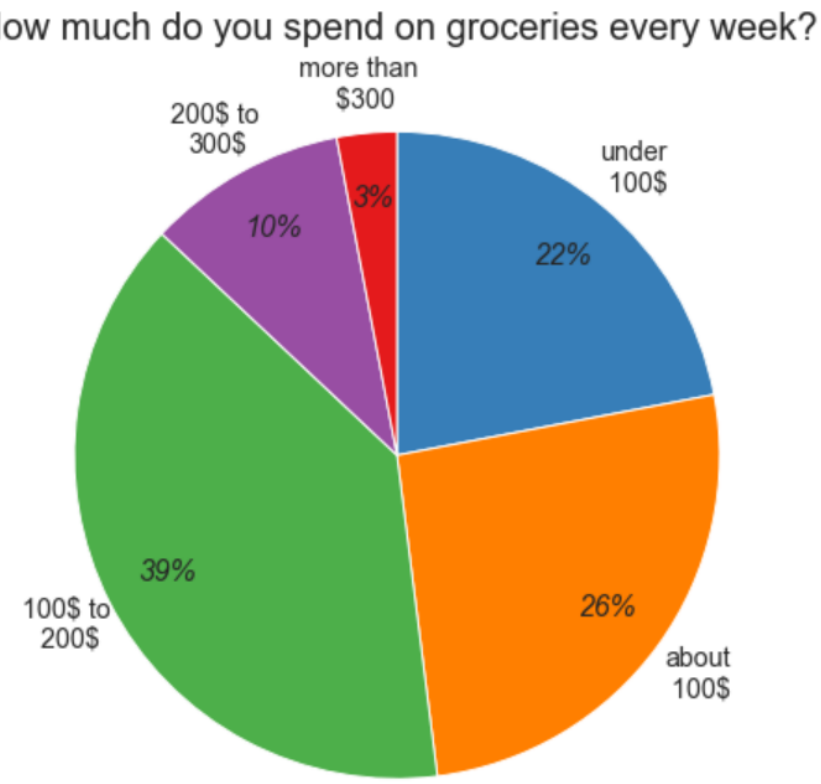


Even though the graph looks meaningless at first glance, I can interpret it as the composition of different expenditures on the groceries, whose percent summation is 100%. As a viewer, the graph seems to be irradiating from the center with equidistant squares. There is ambiguity in this graph because it can be interpreted in two ways:

1st, the area of the squares represents the percentage, including the underneath area covered by the top squares (such as 10% square covered by 3% square)
2nd, only the border of the squares represents the percentage, that is, the 10% is represented by the 10% square area, minus the the area of 3% square area.

In either way, the lie factor is significant. In the first case, the pink square is 9 times larger than the red square, while 10% is only 3.33 times larger than 3%. As such, the lie factor is 2.7. For the second case, the pink border is 8 times larger than the red square, whhich causes a lie factor of 2.4. This effect is even magnified for squares of larger percentages.

Additionally, the graph has unnecessary milk-color background and the grid texture over the squares makes the design prioritized over the figures in this case.
The improved visualization for Figure 2 that I propose is the pie chart as follows:

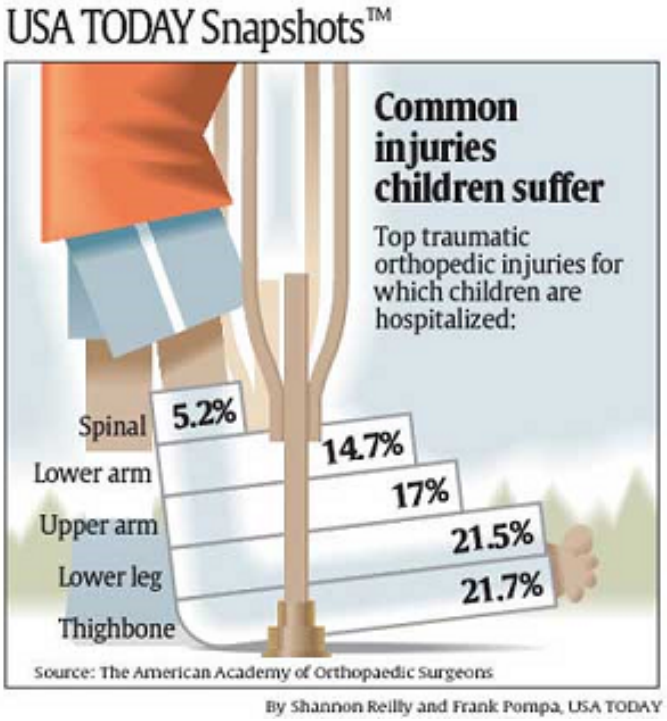


The improvements of this pie chart compared to the above figure:
- No chartjunk (no colored background and grid texture)
- The lie factor is 1, which means there is no exaggeration in the data. Each slice faithfully represents the magnitude of the percentage by area.
- There is no ambiguity in this graph: it can only be interpreted in one way.
- The percent number is lying directly on the slices, which can help viewers track the percents.
- The labels are also next to the slices, indicating the range of groceries expenditure.

Exercise 2 (2 points)

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. Specify precisely the visualization you have selected. If you provide a link, it must be functional and unambiguous (otherwise, you get zero points). It is safer to insert the picture in your report.

I have found this figure on the USA TODAY snapshots website, which tells about the common injuries that children suffer.



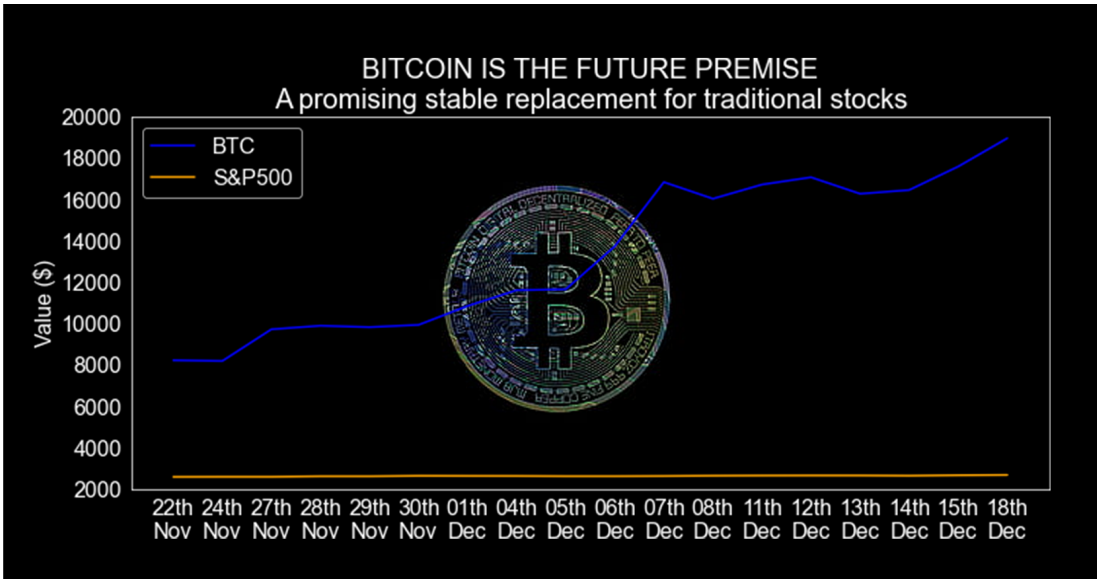
This figure is significantly bad for various reasons:

- First of all, is the redundant chartjunk. This figure focuses mainly on the visual representation for beautiful design instead of showing the figures. Additionally, the background image may look disturbing to some viewers due to the injury illustration. Also, the bar graph is not even completely horizontal but is rotated a little bit, which can confuse viewers to make comparisons between the injuries. The crutch also cuts the bar chart into two halves, resulting in a disrupted graph that viewers cannot measure the scale of the percentages.
- Secondly, the title is "Common injuries children suffer". When viewers look at injuries on horizontal bar charts, viewers have an expression that 5.2% of U.S children have spinal injuries, while the actual context is only revealed at the very end of the subtitle (hospitalized children). This is misleading, as the title should be "Common injuries hospitalized children suffer" instead
- Thirdly, the data-ink ratio is not maximized. At least if the authors want to guarantee the copyright, then the newspaper logo is enough. The source and the authors can be omitted.

Exercise 3 (7 points)

(a) Satoshi is running a crypto business. It is very turbulent with fake media spreading rumors of bubbles and pyramid schemes. Your goal is to help Satoshi convince the public that bitcoin has performed better than the S&P 500. Use the provided data (BTCvsSP500.csv), which contains the daily closing prices in US dollars for the bitcoin and S&P 500 index, respectively, 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.).

This is the figure that I devised for this purpose

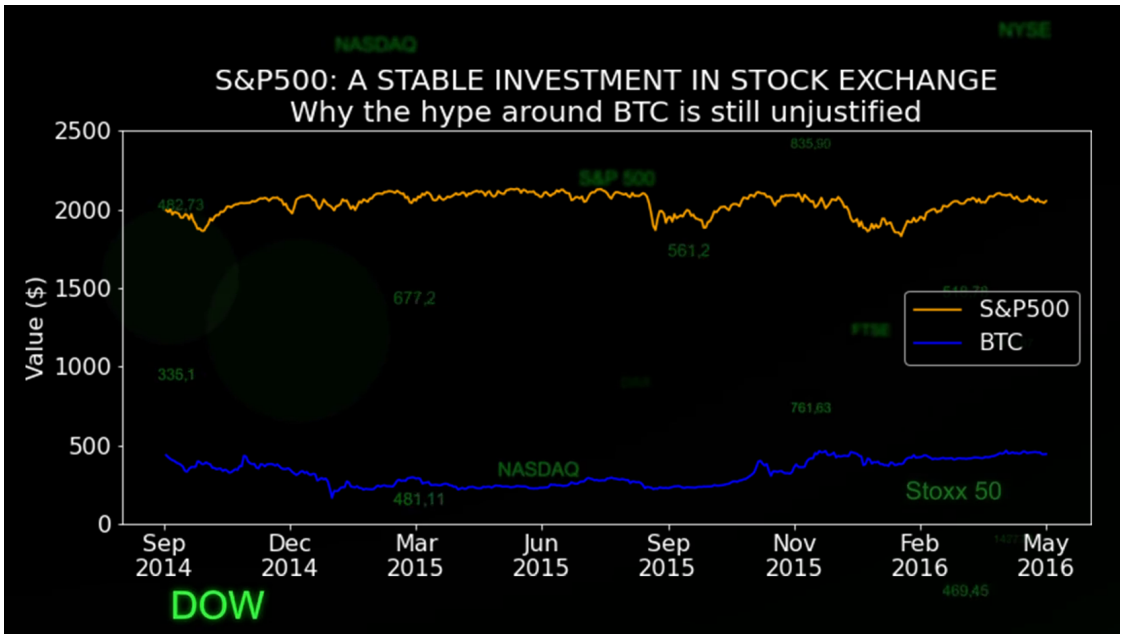


These are the reasons that explain why this graph is misleading and can lead the general public to believe that bitcoin performs better than the S&P500

- The title is misleading, as it tricks viewers into thinking that BTC performs better than the S&P500 stocks and that it is wise to invest in BTC instead of the stocks.
- Chartjunk: a bitcoin figure in the middle tricks viewers into believing BTC is a glistening, ideal investment. Meanwhile, in reality, it is not.
- Cherry-picking the dates: in the true data, both BTC and S&P500 grow up and down, and neither of them has consistently higher values than another. However, I only choose the dates when bitcoin only rises in value and is much higher than the stocks, which stay stagnant throughout the days.
- Time interval is also small (less than 20 days), so it is misleading to generalize the value of BTC over this short time. No year is included, so the viewers do not know this is outdated data. The latest updates may have different trends.
- The y-axis starts from 2000 instead of 0, which causes an exaggeration in the referenced value of BTC concerning S&P500

(b) 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.

This is the figure that I devised for this purpose



These are the reasons that explain why this graph is misleading and can lead the general public into believing that the S&P500 performs better than the bitcoin. I used the same strategies like in part (a), with the reverse effects

- The title is itself misleading, as it tricks viewers into believing that S&P500 stocks performs better than the BTC and it is wise to invest in the stocks instead.
- Chartjunk: the figure is scattered with matrix-style stock brand names with a green font color, which can make the viewers associate positivity with S&P500. In reality, the values of Bitcoin and stocks depend on the current global affairs, so it is untrue that this trend will stay unchanged for many decades.
- Cherry-picking the dates: in the true data, both BTC and S&P500 grow up and down and neither of them has consistently higher values than another. However, I only choose the dates where the stocks is always higher in value than the bitcoin to make an impression that the stocks is more valuable. However, in this period, both of them stay stagnant, suggesting the viewers that the S&P500 has long-term stability in investment returns than the Bitcoin.
- The time interval is long, spanning two years from september 2014 to may 2016. This helps to convince the viewers in the stability of the stocks.

(c) 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.

- For part (a) figure:

The y-axis starts from 2000 instead of 0, which greatly diminished the value of S&P500. For example, let us consider Dec 18th, where the height of BTC is around 26 times that of S&P500. The actual value difference is 18500/2800 = 6.6 times. Therefore, the lie factor in this graph is 26/6.6 = 3.9, which is highly exaggerated.

Conclusion: the plot in part (a) overestimates the relative performance of the two financial instruments by a factor of 3.9

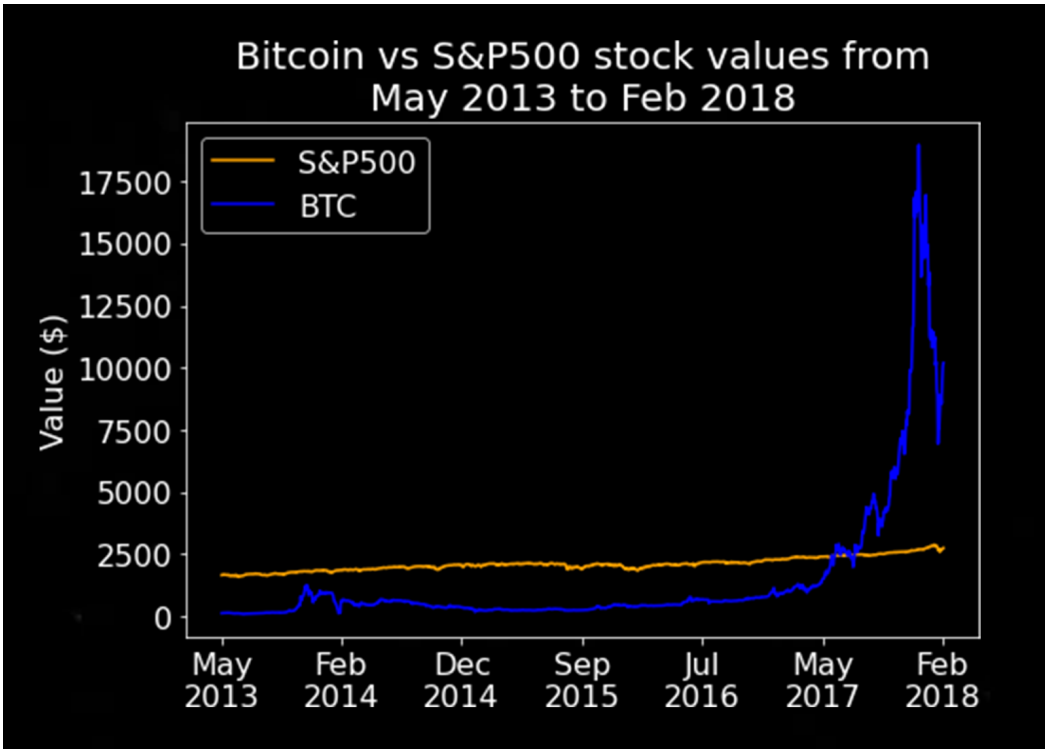
- For part (b) figure:

The lie factor is one this time because the y-axis starts from 0. I do not use this strategy because the BTC value is already very low, so making the y-axis start from 100 does not inflate the lie factor significantly.

Conclusion: the plot in part (b) neither underestimates nor overestimates the relative performance of the two financial instruments

(d) Jorma is a student at Aalto University. He is impartial because he has no money, bitcoins, or S&P 500 ETFs. He decides to start a blog of graphical designs of important topical datasets. Help Jorma and follow the principles of Tufte as closely as possible, and create a plot for the relative performances of the bitcoin and S&P 500. Justify your choices, and describe how/whether you can improve your visualization even more.

This is the figure that impartially demonstrates the data according to the Tufte's principles



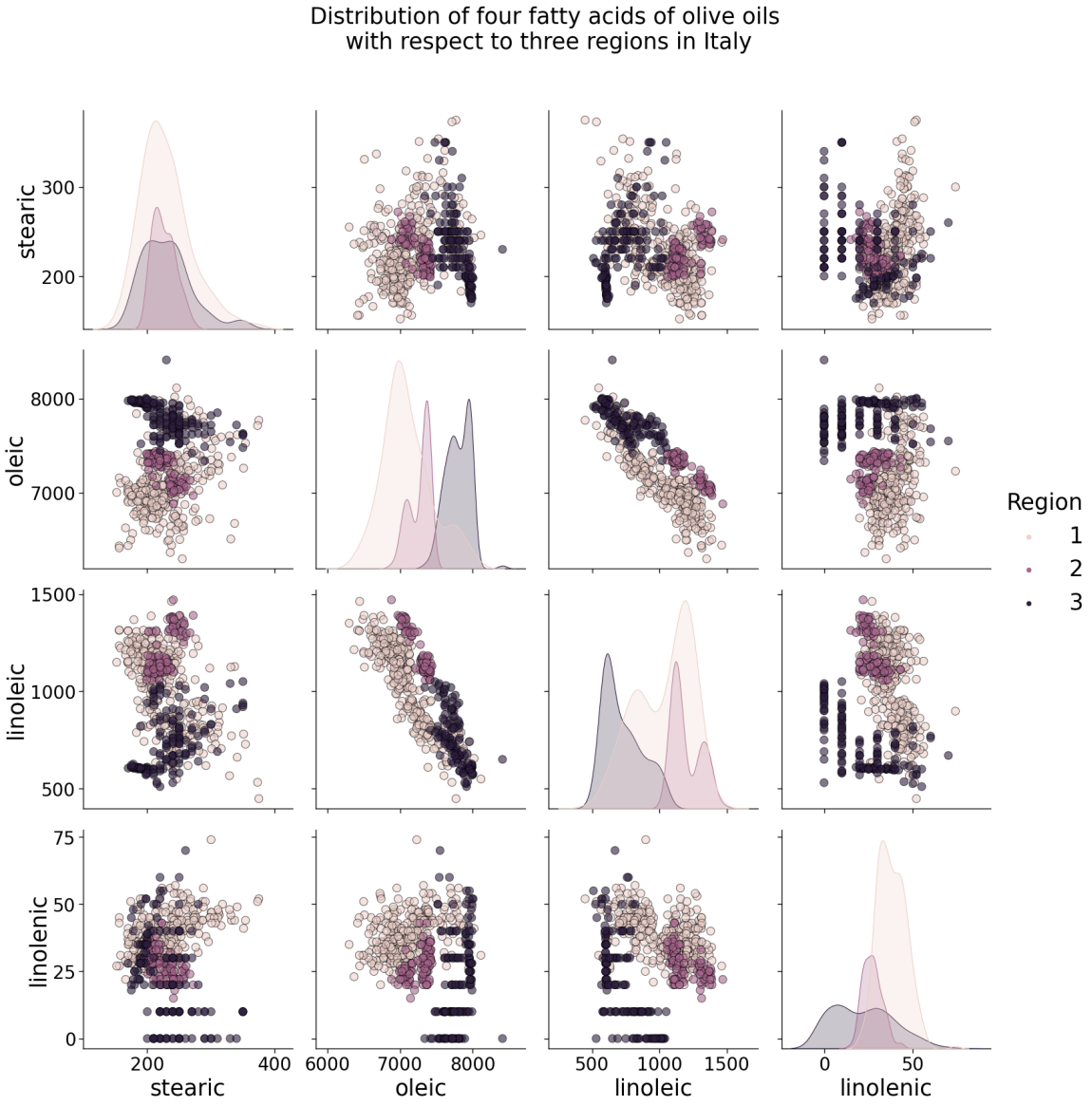
1. Maximizing the data-ink ratio: there is virtually no redundant data-ink in this figure, as removing any components would result in a loss of information. Therefore, the data-ink ratio in this figure is very high.
2. Reducing the chartjunk: this graph has no distracting visuals, such as the bitcoin image or the stock brand names like the previous figures in parts (a) and (b). Additionally, there are no grid lines, grid points, or matching verticle lines along the time xticks.
3. The data density is not too high, which does not make the figure clutter
4. Small multiples: comparisons must be positioned within the eye span for the viewer to make comparisons at a glance. Because the line graph of BTC and S&P500 are plotted against each other, the viewer can track the relative changes in data.
5. Ensuring the aesthetics and techniques: this figure has a golden ratio. It also puts the temporal axis horizontally, which is fit for human eyes to track the changes in BTC and S&P500 over the years.

I think some improvements can be made, such as converting to white background, if some viewers have a problem reading figures in black background. Also, the line weight of the line graphs can be made heavier.

Exercise 4 (3 points)

Visualize the Olive dataset, available at the MyCourses page. This dataset contains 572 olive oil samples from three different regions of Italy. For each sample, the normalized concentrations of eight fatty acids are given. The first variable indicates the row name, the second region, the third the area, and the remaining columns provide the fatty acid concentrations. All numbers are separated by a comma, and the first row gives the column labels. Select at least four features, and create small multiples (trellis), a 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 regions. Try to show the difference between the regions, and maximize the data-ink ratio, within reason.

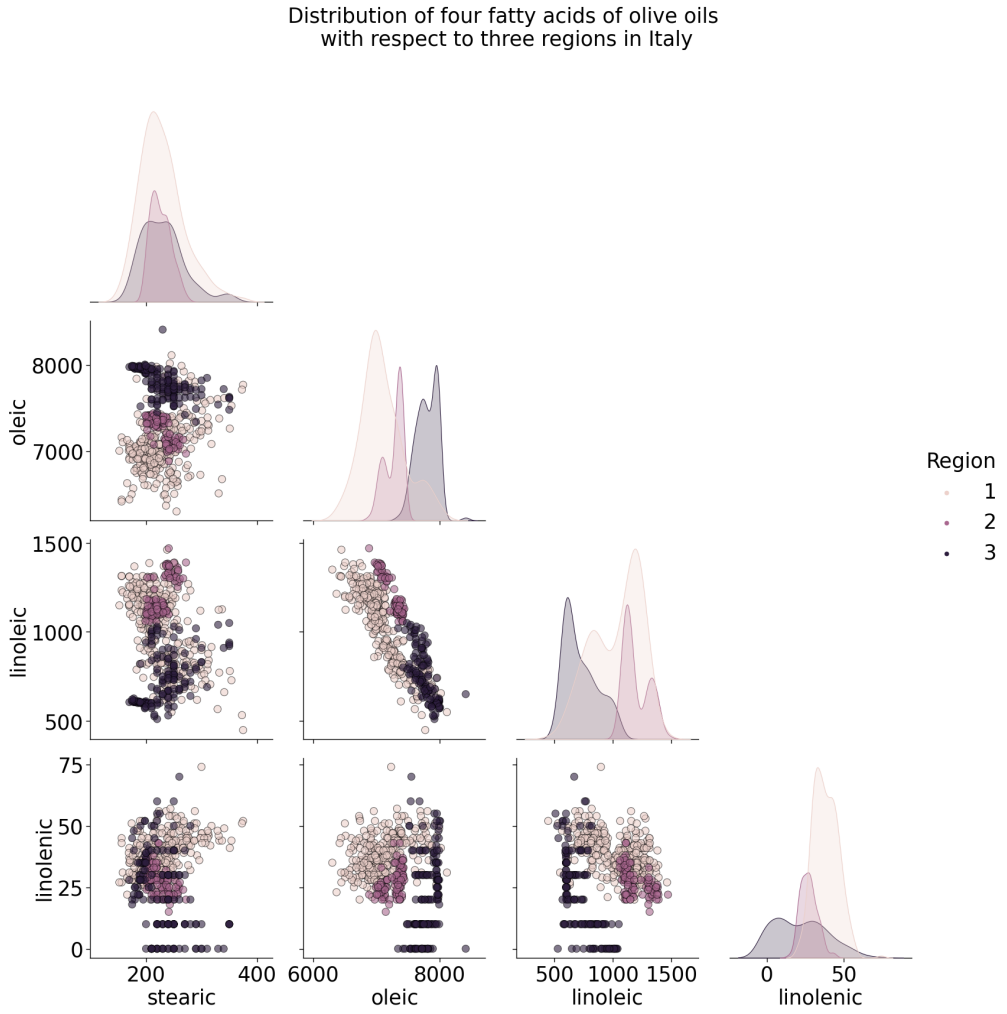
This is the small multiples (trellis) of different visualization for each pair of features, arranged as a matrix. This type of graph is also called a pairplot.



From the figures, we can study different rates of compositions of the olive oils coming from different regions in Italy. Each graph is a scatterplot between each pair of features, while the main diagonal is the kernel density estimate (KDE) of each feature. Each dot in the scatterplot serves as a multifunctioning element, where the coordinate indicates the value of each feature, while the color indicates which region it belongs to. From the legend, the light milk color stands for the first region, the purple color for the second region, and the dark color for the third region. The four features I choose in this exercise are the fatty acids stearic, oleic, linoleic, and linolenic. By looking at the pairplot. For example, for stearic acid, all three regions seem to have the same mean, while for oleic, the oils in region 1 have less oleic than region 2, and the oils in region 2 have less oleic than region 3.

Additionally, the scatterplots can reveal the relationship between each acid. For example, oleic and linoleic are in inverse proportion across all three regions. Regarding the data-ink ratio, where one can see that the upper half corner is just a duplicate transpose of the lower corner. As such, the data-ink ratio in this figure is one-half (0.5). To maximize the data-ink ratio, we can delete the upper corner and keep only the lower corner of the pairplot

This is my suggested improvement to the above graph to maximize the data-ink ratio.



Now with only the lower corner, the scatterplots are no longer repeated for the same pair of features. The data-ink ratio is maximized at 1 without any further redundant data-ink