Student Name : Quynh Diem Luong
Student Number: 808244

Information Visualization

Assignment 1

Exercise 1

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

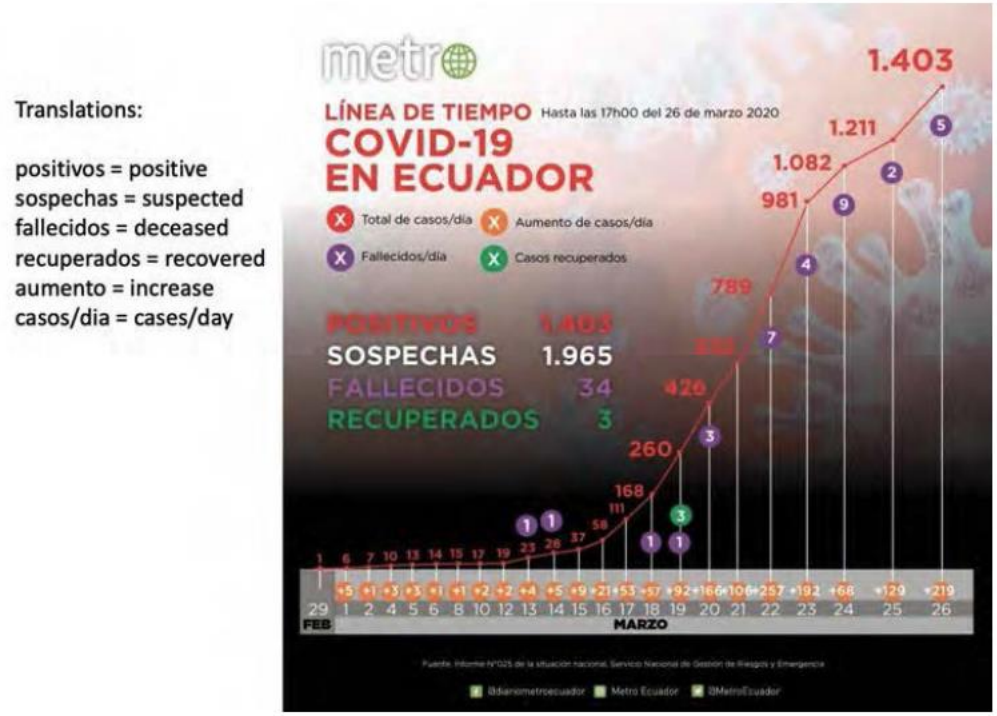

Figure 1. Statistics of early COVID 19 outbreak in Ecuador (March 2020)

Answer:

According to Tufte's principles, there are a total of five keys components in data visualization that ensure a good design principle: Data-ink, Chartjunk, Data density and small multiples, aesthetics and techniques.

We easily see that the visualization in figure 1 contradicts at least four items from Tufte's list.

The first violated principle is **data-ink.** Data-ink contains empty space and ink. It is a non-erasable and non-redundant core inside the graphics. Tufte also states that by erasing data-ink, we reduce the amount of information transmitted by graphic

From the visualization above, we see that there are two main components of the visualization which are

- A table listing the total number of people for each COVID status category (positive, suspected, deceased, and recover)
- A line graph that shows the total number of people who tested positive for COVID. For each data point, the graph also includes the number of people who is being suspected, deceased and recovered.

We easily see that those two pieces of information are overlapping on each other. In other words, if one of the graphs is being removed, we can still derive the necessary information from the remaining one. Thus, the amount of information transmitted by graphics stays the same if we erase data-ink. It is clear that the data-ink principle does not hold for this figure.

The second principle that does not hold in figure 1 is **chartjunk.** Tufte's principle states that chartjunk is the interior decoration of a graphic that does not tell the viewer anything new. The purpose of chartjunk could be to decorate the graphics, make data appear more lively, or make graphics appear more scientific and precise. The visualization includes an eye-catching coronavirus in the background. This background does not transmit any new information to the viewer. As a result, eliminating it has no impact on the goal of communicating early COVID statistics in Ecuador. It goes against Tufte's notion because in this situation, it's more about showing the graphic ability of the designer than providing the facts.

The third principle that the visualization violates is the **data density and small multiples.** Our eye can only distinguish patterns of about 40 cycles to 150 cycles per degree in computer graphics. We see that the line graph is highly proportional due to a large gap between data points (from 1 to 1403). Thus, we barely see anything on the graph for at least 10 first data points. Due to the steeper slope, the chart gives the impression that the 22-23 day interval has more patients than the 25-26 day interval. However, the rise does not match the numerical representation (which also violates **data integrity** principle). Besides, the x-axis is also too dense that we hardly see the x value in the line graph.

Fourthly, the visualization does not hold the **aesthetic and technique** principle. In this principle, Tufte stated that graphics should usually have a greater length than height because detecting deviations from the horizon is a skill that our eyes have acquired. In other words, it is easier for us to read graphics horizontally. Longer horizontal helps us emphasize the cause variable. However, we see that the visualization was presented in a ratio 1:1 (which is a square). This ratio is far from the suggested ratio 1:1.618. Thus, it violates the graphics proportion rule that Tufte has set.

(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.

Answer:

Figure 2 below shows an original visualization (left) and the improved visualization (right). In the improved version, I visualize those data under a pie chart.

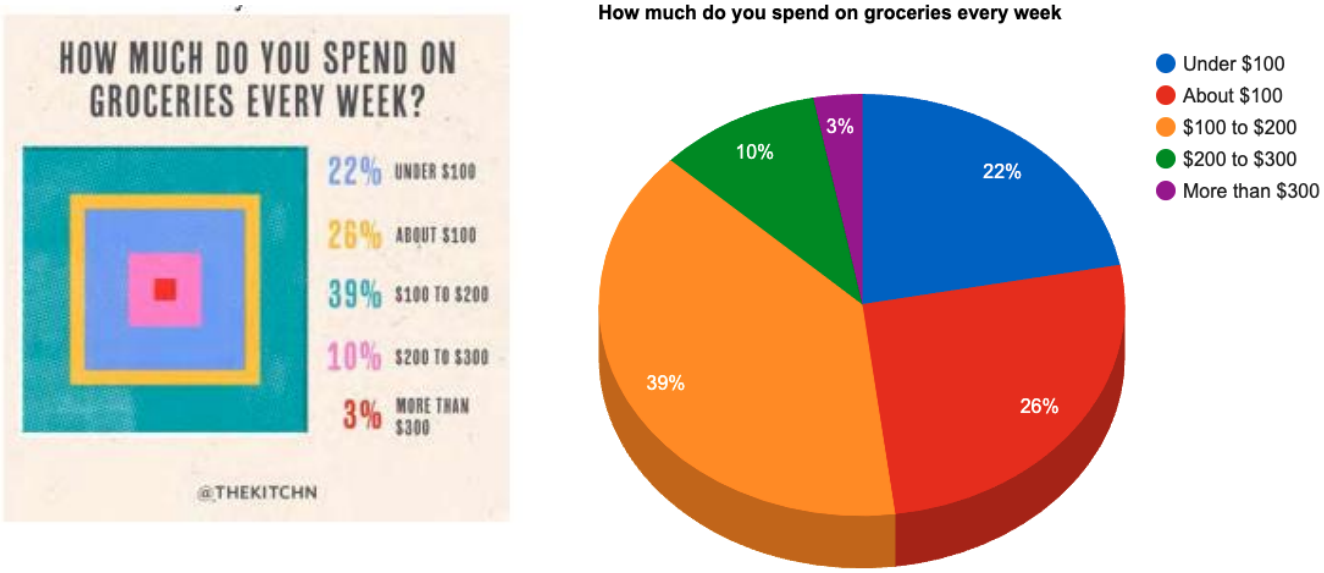

Figure 2. Original Visualization (left) and New Visualization (right)

One of the main reasons why I decided to illustrate the data under pie chart visualization is because the total amount of percentages for each label is 22% + 26% + 39% + 10% =100%. It is intuitive for the viewer that each arc length represents a proportion of each category while the full circle represents the total sum of data, equal to 100%.

In addition to that, the pie chart helps show proportions and percentages between each category by the size inside the circle. It helps viewers quickly get some important insights such as which category is the most popular one, which one is the least popular, and how each category is proportional to the others. The original visualization causes some confusion for viewers because 26% looks smaller than 22%. Besides, the proportion does not reflect each value correctly. For instance, we all know that 3% is approximately 1/3 of 10%. However, in the original visualization, it looks like 3% is 1/9 of 10% if we make a simple measurement by eyes. This violates Tufte's data integrity principle. By using a pie chart, we can obey this principle in a simple and easy way. Also, viewers can see a data comparison and understand the information immediately.

Exercise 2

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.

Answer:

Figure 3 shows a good example of a visualization that follows Tufte's design rules. The visualization shows some statistics of fiction book sales in bar chart format.

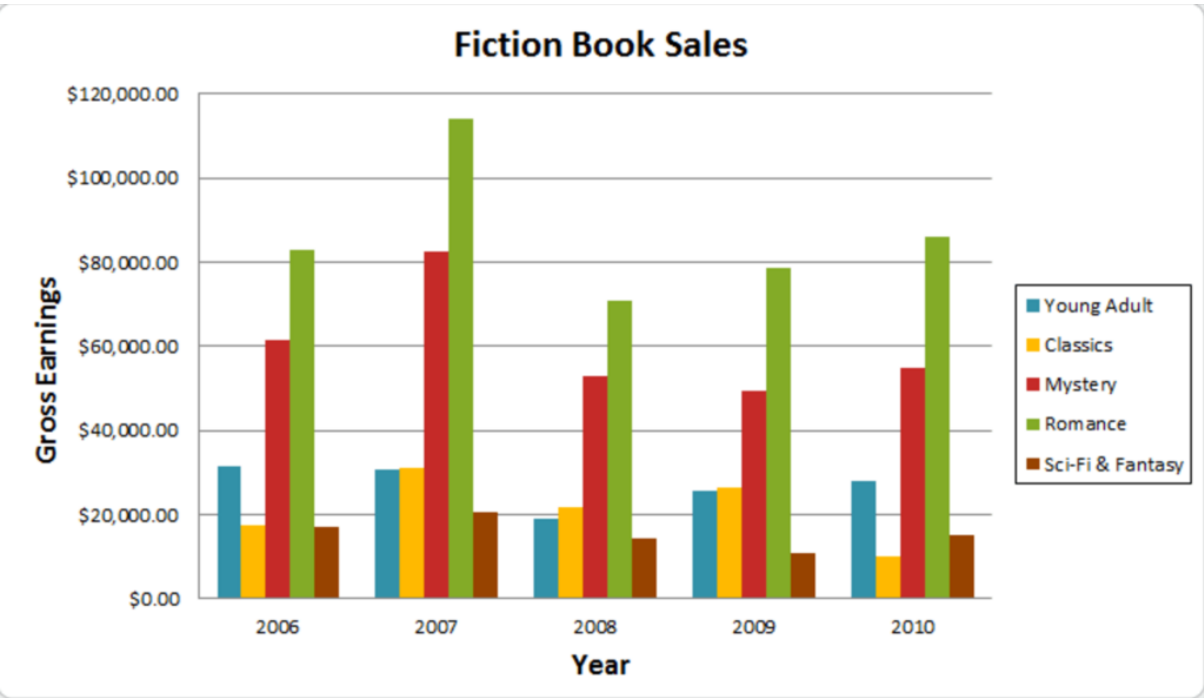

Figure 3. Fiction Book Sales in Different Group Age

At a first glance, viewers can quickly get some important information such as categories (young adult, classics, mystery, romance, sci-fi and fantasy), year (from 2016 to 2010), and gross earnings for each category.

The first impression we can grasp from this visualization is the proportion of graphics. It is extended horizontally in a reasonable ratio. Thus, the causal variables and labels are eye-catching as well as easy to read. It obeys Tufte's rule of a **good graphics proportion**.

Secondly, the **data-ink** for this visualization is highly insightful and non-erasable. If we erase any data-ink (label, variable, column, etc) contained inside this visualization, the amount of information transmitted by graphics is highly reduced. Every bit of ink on visualization presents

new information and has a reasonable interest to the viewer. Thus, it holds a good Tufte's principle of data-ink

Thirdly, the chart does not have any **chart junk** because each of the visualizations is used to depict the illustration of fiction book sales. There is no unnecessary decoration contained in the figure just to make the data appear more lively.

Exercise 3

(a) *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 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.*

Answer:

The given visualization shows that Bitcoin has performed better than the S&P 500. The line graph was conducted using the matplotlib library in Python.

Figure 4. Bitcoins perform better than S&P500

In fact, by using a portion of the provided data, it is sufficient to show that Bitcoin has outperformed the S&P500. I selected the sub data that only included the price in 2017 (duration from August 2017 to January 2018). In this duration, Bitcoin has been famous for its widely optimistic price and

everyone seems to invest in it. Thus, the price was pushed really high compared to S&P500. This visualization creates an illusion that Bitcoin has maintained a high price all the time and the trend will keep going on. Meanwhile, the S&P 500 seems to remain stable. Thus, it would be far better to invest in Bitcoin rather than S&P 500.

(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.

Answer:

Figure 3 shows a visualization that S&P 500 performs better than Bitcoin. This line graph also uses the same trick as what we have done in the previous question. Instead of plotting the whole dataset, we only visualize a subset of data in the year of 2015. Bitcoin did not have much monetary worth at the time of 2015 and most people were unaware of its existence. We easily see that S&P 500 price is much higher than Bitcoin. If Warren shows this chart to the public, they would clearly agree that S&P 500 is more worthy than Bitcoin.

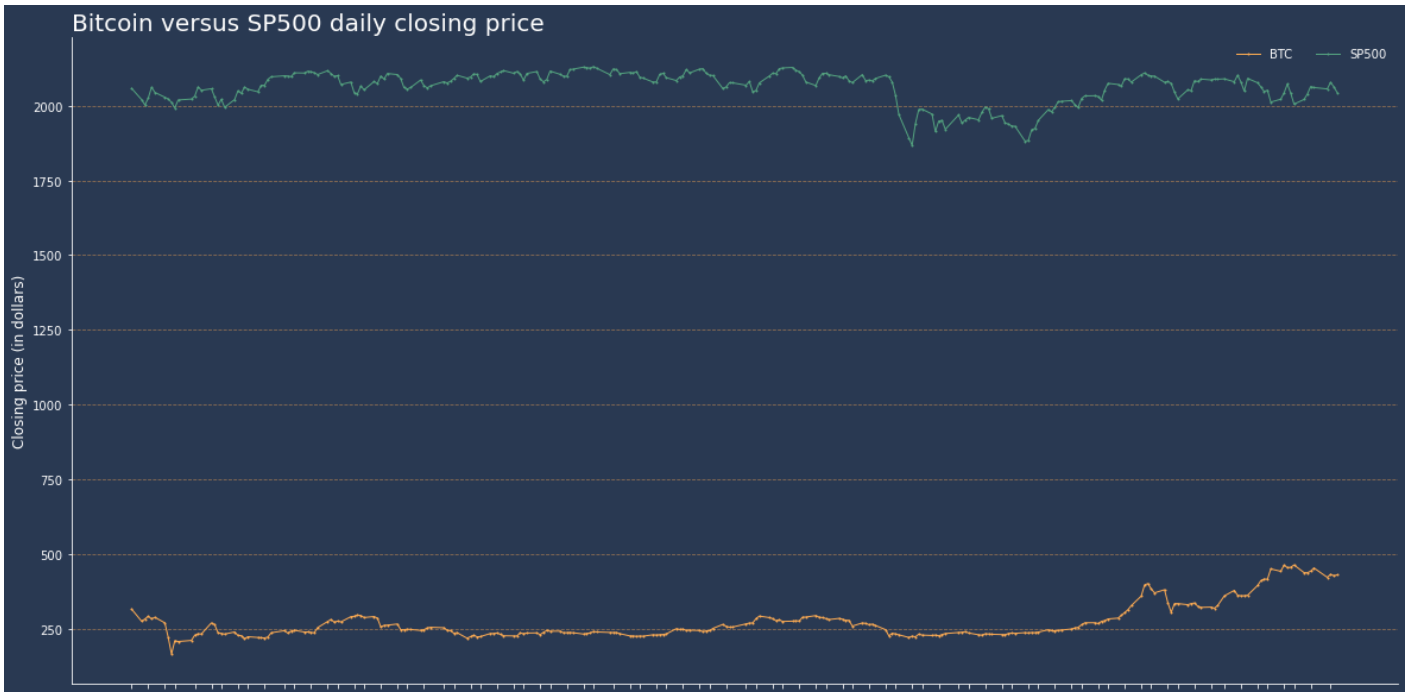

Figure 5. S&P500 performs better than Bitcoin

(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

Answer:

Both of the above plots are created using only a part of the dataset, and there are no other lying tricks. As a result, the numerical changes are identical to the impact depicted in the graph. Its goal

is to persuade or deceive the public into believing certain assertions, such as which one performed better. Thus, the size of effect shown in graphic is equal to the actual effect in data. As a result, the **lie factor is 1**

(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

Answer:

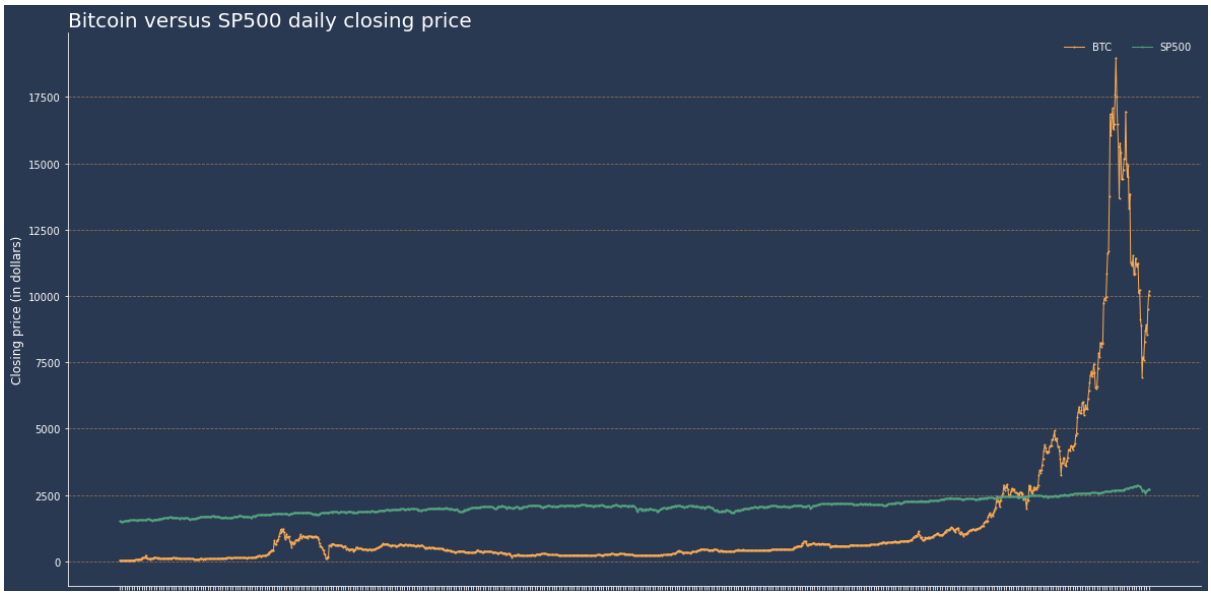

Figure 6 Bitcoin versus SP 500 daily closing price

Figure 6 shows the proposed visualization for Jorma's case. In this visualization, we use all the provided data to create this graph. When it comes to data integrity, the graph provides an overview of Bitcoin and S&P 500 prices for a lifetime. Meanwhile, in the two previous figures, we only provided the audience with "half of the truth".

The proposed chart follows Tufte's example in several key factors such as data-ink, chart-junk, and graphical integrity. We easily see that the line graph has a high **data-ink** ratio. The data-ink is also a non-erasable and non-redundant component of the whole graphic. By erasing one of them, we will reduce the amount of transmitted information inside the graphics. Secondly, there is no **chart junk** used as interior decoration for the graph, each component has a statistical meaning to

deliver to the audience. Thirdly, the data is **well proportioned**. The prices of both currencies are clearly depicted, which helps audiences to get some insights into where they should invest in.

However, there are also some **improvements** that we can see from the graph. We easily see that each data point in the graph was visualized regarding a specific timestamp. However, we did not display those timestamps on the x-axis. The x-axis values seem unknown to the audience unless someone explains that to them. Since displaying all the timestamps on the x-axis could be too dense, a proposed solution is that we can only show the start and end date on this given axis. By indicating the start date and end date for each duration, the audience will be given more insights into when the dataset is conducted. Also, the label and number could be bigger so that audience can easily capture the features.

Exercise 4

Visualize the Olive dataset, available at the Mycourses page. Indicate with different colors the three regions. Try to show the difference between the regions, and maximize the data-ink ratio, within reason.

Answer:

Figure 7 below shows the scatter plot with four features which are palmitic, palmitoleic, stearic, and oleic. The visualization was conducted using the Python seaborn library.

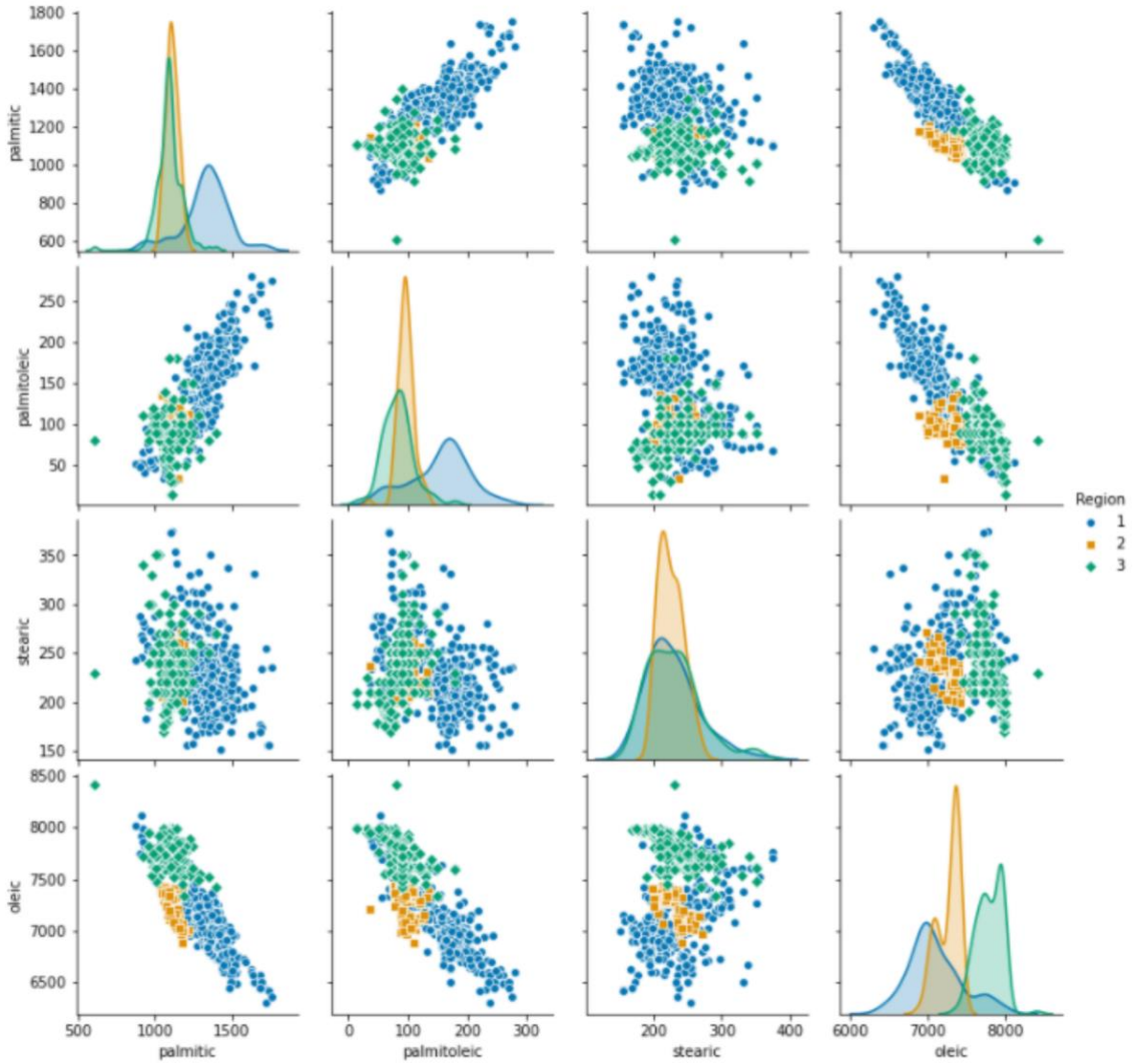

Figure 7. A scatter plot with four features

In this graph, we easily see **different colors for the three regions**. Each region was represented in a distinctive color (i.e. blue for region 1, orange for region 2, and green for region 3). If we take a closer look at the graph, we can see that **different markers** are also used to represent each region (circle for region 1, square for region 2, and diamond for region 3). Those techniques help reduce clusters in the scatter plot. Even though the cluster contains a lot of overlapping points, our eyes are still able to figure out how each region's olives are distributed in each subplot based on their colors and shapes. As a result, it helps the audience easily distinguish data for different regions and the interpretation of the figure is clearly improved.

Consider that a good visualization should be able to accommodate multiple types of viewers if feasible, in this instance, the seaborn library contains a collection of color-blind colors that aid in improving the depiction of the graph from the perspective of a color-blind person. By using

clear and transparent representation to distinguish the difference between regions, we also help maximize the data-ink ratio.

However, we realize that there are still some erasable data-ink inside our chart. If we consider our left diagonal figures as a pivot line, all other plots are identically symmetric to this line. In another word, left upper corner plots are the reflections of lower right corner plots. If we flip all the plots from the upper left corner through the pivot line, we will get identically right corner plots. Thus, we can remove one of them as those identical plots do not yield any new information (called redundant data-ink). The improved scatter plot is shown in figure 8 below

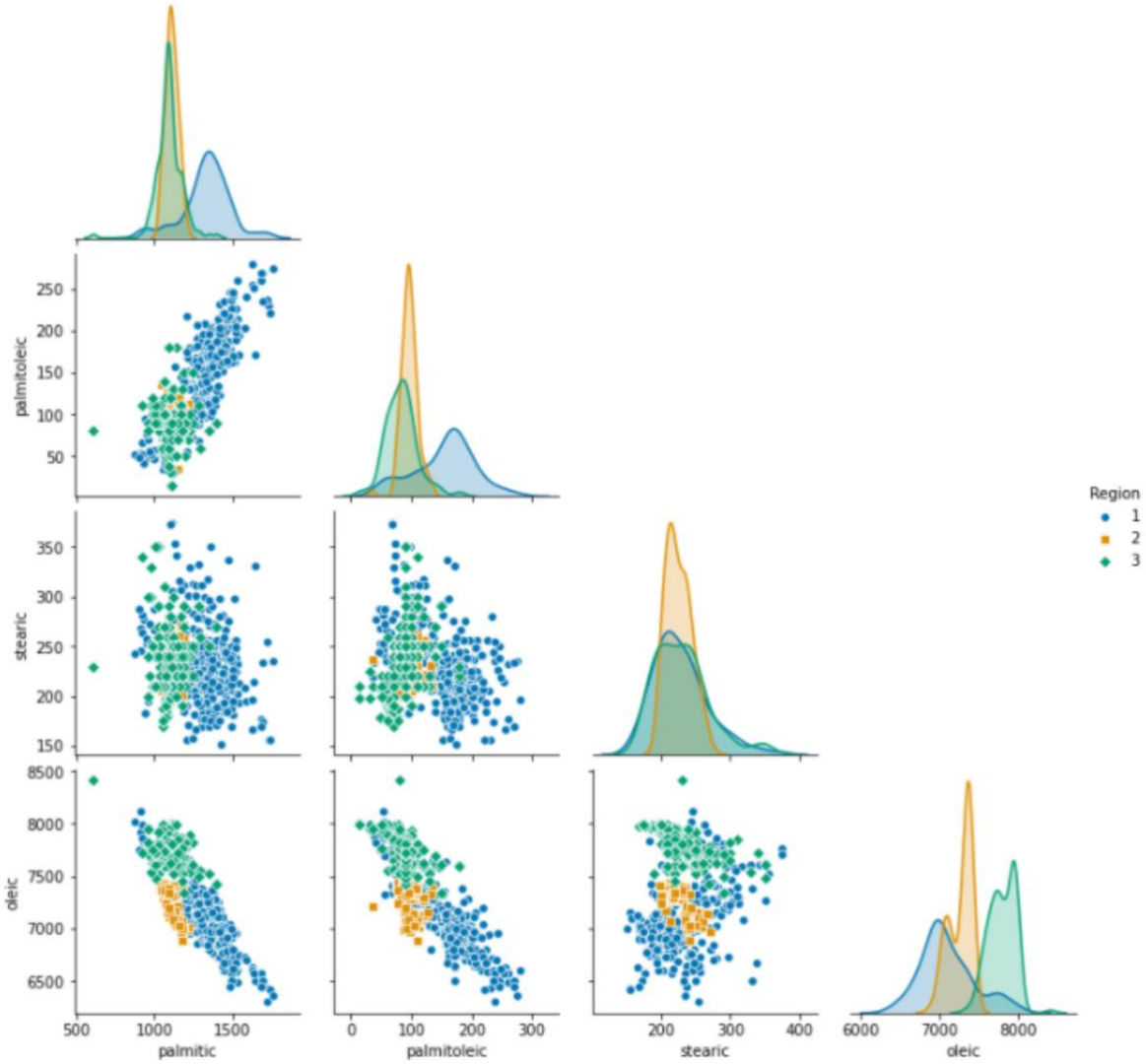

Figure 8. Improved scatter plot

We see that the **data-ink ratio is maximized** inside the new visualization. Each plot is now identical and represents new information for audiences.