Name: Quynh Diem Luong
Student Number: 808244

Assignment 2

Exercise 1

Take a look at the EU Open Data Portal database. Write a short (2–3 pages) report that describes your approach, obtained results, and the methods you used. Explore at least three of the topics discussed in the course (see below) and explain in your report how you have used them. The purpose is not to do an analysis that covers the data from all possible angles. Focus on one or two aspects and make a clear visualisation of them. Try to relate your work to the topics discussed in the course, i.e., Tufte's principles, pre-attentive features, Gestalt laws, etc., when preparing the visualisations. Also, make sure your images convey at least some of their information when printed on a black and white laser printer.

Answer

In this exercise, I chose COVID19 testing dataset provided by European Centre for Disease Prevention and Control. The dataset consists of 13528 rows and 12 columns.

1. Methods

In this exercise, we use Python and its built-in libraries (pandas, matplotlib, seaborn). Pandas will be used to preprocess and filter the raw data. In addition, we use matplotlib and seaborn to visualize the dataset.

2. Approach

COVID19 testing dataset was conducted to measure five different features (new cases, the number of tests done, population, testing rate, positivity rate) weekly in European countries. The dataset can be shown in figure 1 below

	country	country_code	year_week	level	region	region_name	new_cases	tests_done	population	testing_rate	positivity_rate	testing_data_source
0	Austria	AT	2020-W15	national	AT	Austria	1832	12339	8901064.0	138.623877	14.847232	Manual webscraping
1	Austria	AT	2020-W16	national	AT	Austria	684	58488	8901064.0	657.089984	1.169471	Manual webscraping
2	Austria	AT	2020-W17	national	AT	Austria	447	33443	8901064.0	375.719128	1.336603	Manual webscraping
3	Austria	AT	2020-W18	national	AT	Austria	311	26598	8901064.0	298.818209	1.169261	Country website
4	Austria	AT	2020-W19	national	AT	Austria	263	42153	8901064.0	473.572598	0.623918	Country website
...	...	...	...	...	...	...	...	...	...	...	...	...
13523	Sweden	SE	2022-W11	national	SE	Sweden	8500	37899	10327589.0	366.968515	22.428032	TESSy
13524	Sweden	SE	2022-W12	national	SE	Sweden	7039	35079	10327589.0	339.663013	20.066136	TESSy
13525	Sweden	SE	2022-W13	national	SE	Sweden	4956	27521	10327589.0	266.480395	18.008067	TESSy
13526	Sweden	SE	2022-W14	national	SE	Sweden	4091	26201	10327589.0	253.699097	15.613908	TESSy
13527	Sweden	SE	2022-W15	national	SE	Sweden	3073	21588	10327589.0	209.032331	14.234760	TESSy

13528 rows × 12 columns

Figure 1. COVID 19 testing dataset outlook

Since the purpose of this exercise is to do an analysis that covers the data from one or two possible angles instead all of those features, thus, we will focus on the top 5 countries that were being most affected by COVID19. The goal of those visualizations is to provide the audience

with the impact of this pandemic. One of the statistics that can show the impact is the ratio of new cases and the population as a whole.

Thus, the next step is to combine the data and sort them in descending order. We see from the raw data that different rows present the static of a country. The first thing we need to do is to combine them in the same place. In this exercise, we use Python with the help of the pandas library to do the data processing. Since the data frame is quite long, figure 2 below will only capture some examples of them. We got this new data frame after combining those statistics into the same place.

Out[71]:

	country	population	new_cases	tests_done	testing_rate	positivity_rate
0	Austria	8901064.0	7825881	515675278	177264.346419	38.951187
1	Belgium	11522440.0	7290908	59052759	8106.338876	45.428700
2	Bulgaria	6951482.0	1146514	9658801	4163.126654	40.082018
3	Croatia	4058165.0	2033443	15785571	61511.231952	36.265954
4	Cyprus	888005.0	463837	22933853	67002.325437	58.263305
5	Czechia	10693939.0	3879059	53475215	14956.948978	31.250232
6	Denmark	5822763.0	5682031	243531361	77550.472280	24.343889
7	Estonia	1328976.0	552259	3187836	6540.148204	75.249932
8	Finland	5525292.0	1884620	19802029	24698.163996	92.682927

Figure 2. New data frame after the combination

Afterward, we use the same library to sort those values in descending order. As our goal is to see the top 5 countries that are most affected by COVID19, we have the final data frame as can be seen in figure 3

	country	population	new_cases	tests_done	testing_rate	positivity_rate
9	France	67320216.0	52609456	454836744	17280.328402	49.536575
15	Italy	59641488.0	31090939	402233867	226805.703549	45.684617
10	Germany	83166711.0	21788122	122332384	3149.486097	67.845720
28	Spain	47332614.0	21630911	160232609	11728.300463	101.168960
23	Poland	37958138.0	12098567	72043882	3569.570607	165.655959

Figure 3. Data after being processed

This is the final processed data frame and we will use it for our further visualizations. We see from the final data frame that the top five affected countries are France, Italy, Germany, Spain, and Poland. In the next section, we will create some graphical visualizations based on this final data. Our goal is to apply three approaches that we learned from the course such as Tufte's principle, Pre-attentive Features, and Gestalt's Laws.

Tufte's principle ensures the visualization follows five main principles (data-ink, chartjunk, data density and small multiples, aesthetics and techniques). By following those principles, we ensure a good design for the audience. Pre-attentive features guarantee users use appropriate colors and hues to make the viewer's interpretation of data getting smoother. Besides, it also makes sure that the graphic is color-blind friendly by avoiding using confused colors. Gestalt's law studies different attributes with specific attention to three factors of a visualization: symmetry, proximity, and closure.

3. Obtained results

This section provides two figures that show the impact of COVID19 on European countries. The first visualization is a bar chart that illustrates the number of new COVID19 cases in the top five affected countries. The second visualization is a pair plot that shows small multiples of those five countries.

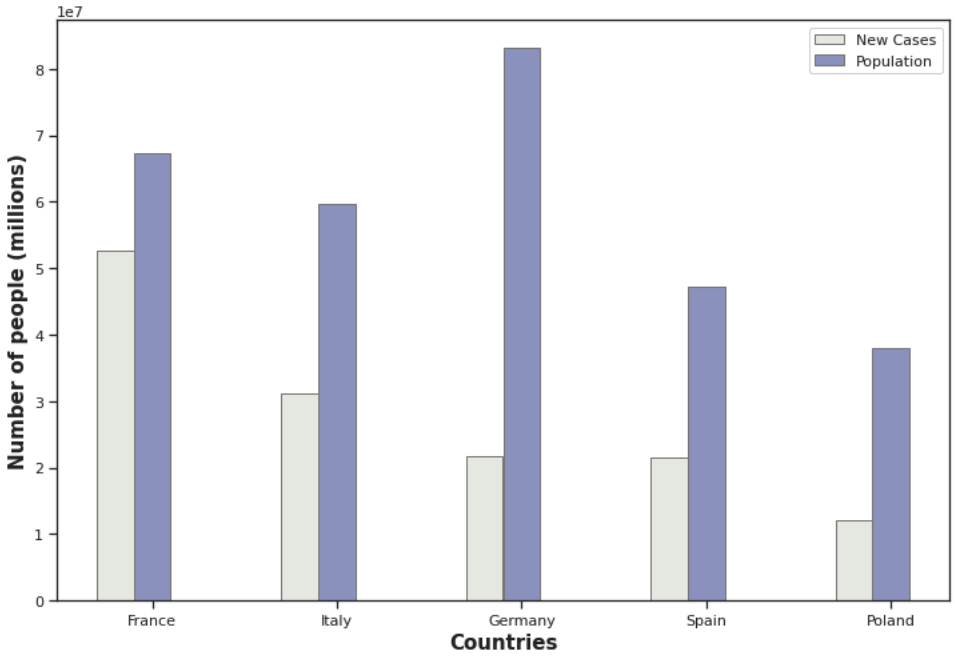

Figure 4. Total new cases in the top five most affected countries

Figure 4 shows a comparison between five European countries regarding the new COVID cases. There are six fundamental principles that Tufte suggests in a visualization: show comparisons, show casualty, data integrity, minimal chart-junk, and a good ratio of data-ink.

- **Show comparisons:** By using the bar chart, the design clearly shows a comparison between countries. Viewers can easily see the number of new cases as well as the population in a specific country. The new cases are highest in France and the ratio between new cases and the total population is also high compared to other countries. Meanwhile, Poland seems to be less affected due to its low density of population.
- **Chart junk:** The chart does not use any extra graphics just for decoration. All the visualization deliver a statistical meaning behind them. Thus, it obeys Tufte's rules of chart junk.
- **Data Integrity:** The chart is extended horizontally in a reasonable ratio. It shows proportions between countries based on real numbers provided by the dataset. There is no scaling applied in order to make the data look more lively.

By avoiding the range of red-green-yellow from the color scale, the visualization also helps color-blind people perceive information easier. The chart color was taken from the suggested pallete for color-blind people.

Figure 6 below shows the same chart after being printed on a black and white laser printer. We see that the visualization is still clear and able to show the difference between represented features. The colors are not too contrasted, our eye can still figure out different data based on the color tone.

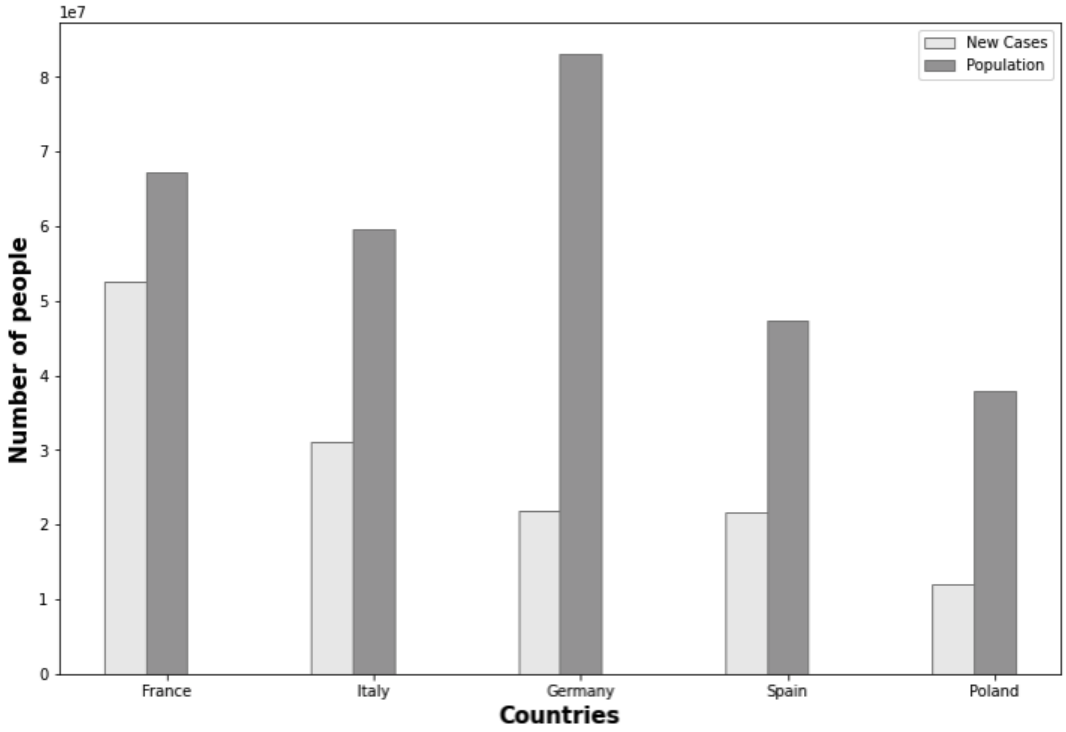

Figure 5. Bar Chart after Black and White laser printer

Figure 6 below shows the pair plot of different COVID19 testing features in those five most affected countries. It differs from the above bar chart in the way that it visualizes small multiples of each country and shows the relationship between those multiples inside the dataset.

Based on Gestalt's law, we see that the visualization has followed the rules of visual displays

- **Proximity:** Our eyes can perceive the rows and columns easily since each row and each column represents different variables. By observing the x-axis and y-axis, we can spot out the relationship between those two variables and determine the given features.
- **Symmetry:** In terms of the diagonal, the trellis is symmetric. All of the graphs are pairwise reflections that differ only in rotation. It makes pattern comparisons easy for the viewers.
- **Similarity:** Each country in the pair plot is represented as a small dot with different colors. In the other words, we group similar objects together using color. Thus, we obey Gestalt laws of similarity.

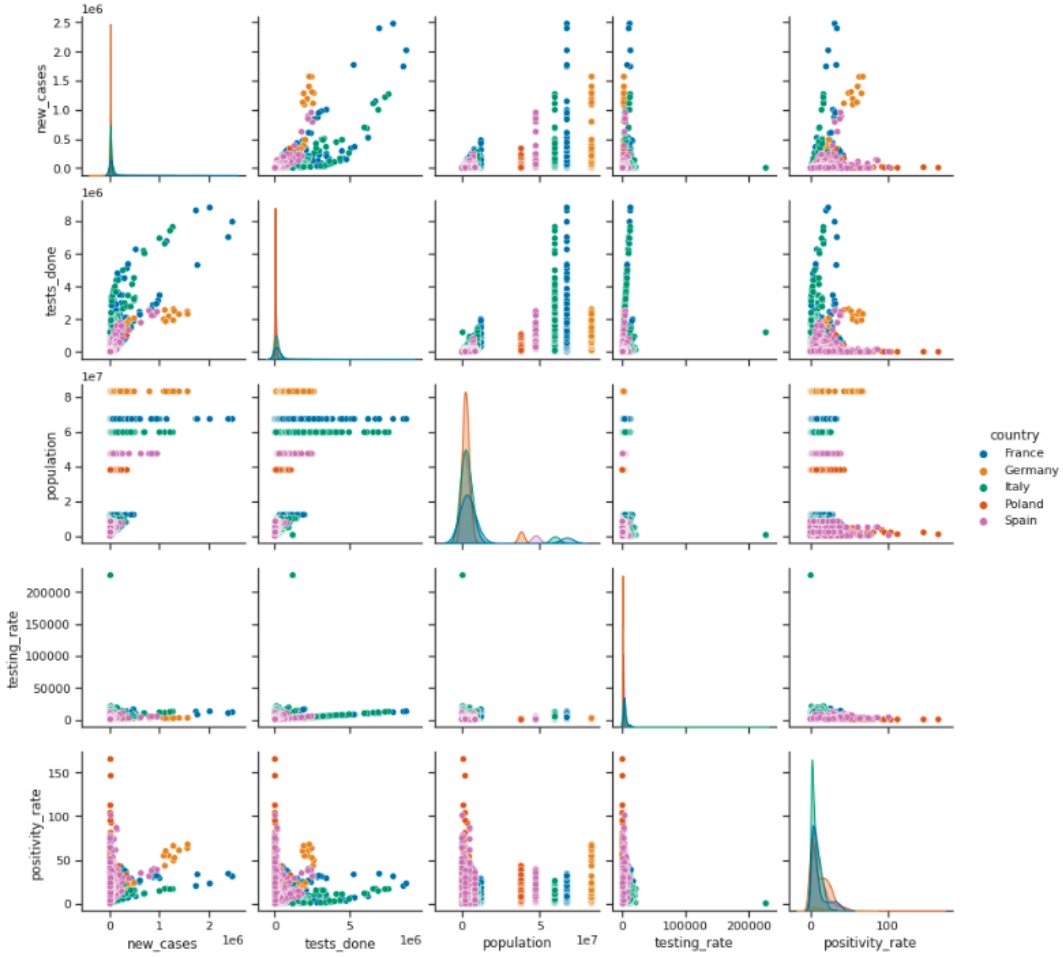

Figure 6. Small Multiples of the top 5 affected countries

Similar to the previous bar chart, the pair plot color scale is also taken from colorblind pallets. Since the chosen colors are designed for color-blind people, it also obeys the pre-attentive features.

Figure 7 below shows how the pair plot looks like after going through a black and white laser printer. Even though we have five different features, we can still spot it out based on its tone. However, our eyes could take a longer time to process the information compared to the bar chart because the grayscale is more diverse in this visualization.

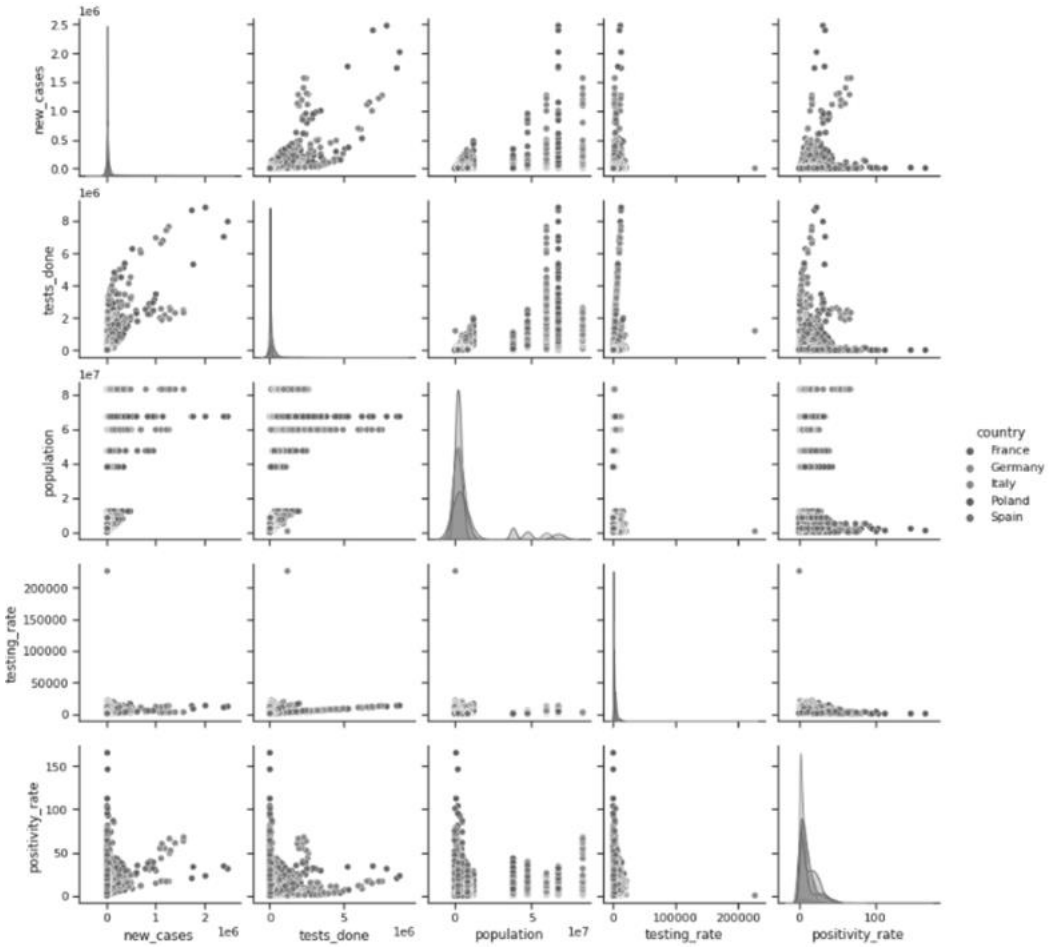

Figure 5. Pairplot through BNW laser printer

Exercise 2

Discuss the simultaneous brightness and contrast effect shown in Fig. 8 and explain it using the Difference of the Gaussians model. Your answer does not have to contain any mathematical calculations. Instead, try to give a clear explanation of why the rectangles appear to be of a different gray tone while they in fact are all of the same "colors." What are the implications of this phenomenon for the design of color scales?

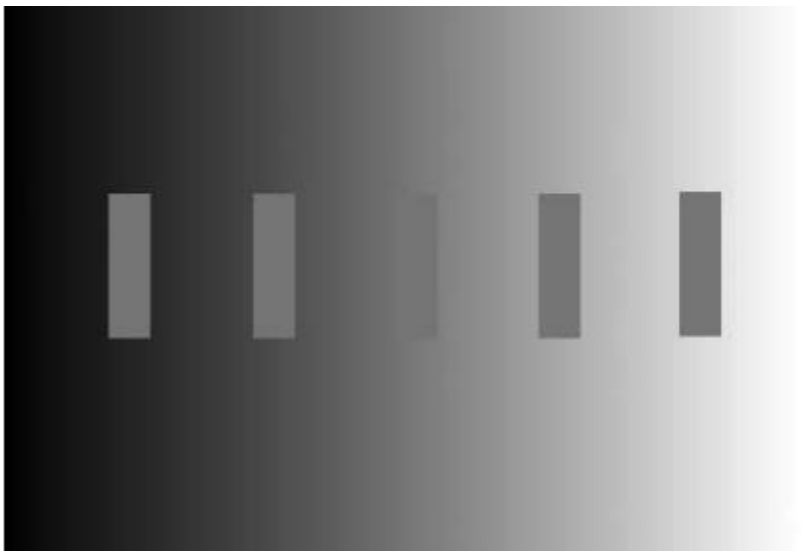

Figure 6. Simultaneous brightness and contrast effect

Answer:

In figure 8 above, we see that all the five patches have the same tone of gray. However, a gray patch placed on a dark background looks brighter than the patch itself on a light background. This illusion is caused due to the way our eyes perceive visual information. According to the Difference of Gaussians, our retina ganglion cells are contained by circular receptive fields. Thus, if light falls at the center of this receptive field, it emits excitation. Meanwhile, if light falls off-center of the receptive field, it emits lateral inhibition. This mechanism explains the difference between physical luminance and how our eyes perceive brightness.

When the gray patch is placed on the dark background, the patch excites ganglion cells more than the background. As a result, the patch looks brighter to our eyes. When the same patch is placed on a lighter background, the background excites ganglion cells more than the patch. Thus, the patch appears darker to our eyes.

From this phenomenon, we see the importance of the color scale in our design. The contrast effect plays an important role in how our eyes perceive simultaneous brightness. By understanding this rule, we can enhance the choice of color scale in our design.

Exercise 3

Fig. 9 is a map of Europe. If you were to show statistical data for comparing the countries, which scale (Greys, Spectral, or YlGnBu) would you use? Justify your answer with at least two reasons based on the functioning of the human eye.

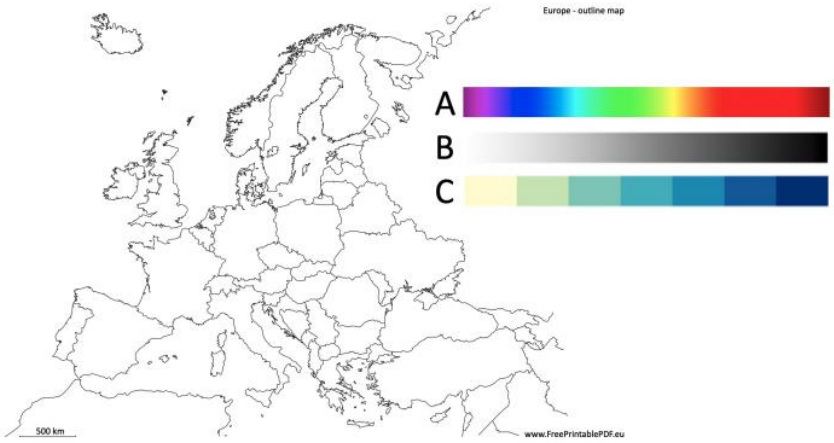

Figure 7. Three color scale and map of Europe with country borders

Answer

In figure 9, to show statistical data for comparing countries, the color scale **YlGnBu** (Color Scale C) would be the best one based on the functioning of the human eye. When our eye interacts with a visual image, the image will be received and processed by V1 (primary visual cortex) and V2 (secondary visual cortex). Those visual cortexes make up to 40% of visual processing and turn the properties into form, color, stereoscopic depth, and motion. When it comes to colors, cone signals are transformed into three types of colors that are arranged perceptually along 3 axes: black-white, red-green, and yellow-blue. Since YlGnBu (Yellow, Green, Blue) scale contains colors that are distinct from each other, it helps viewers to immediately recognize the difference when comparing the countries.

Meanwhile, the Color Scale A contains a wide range of distinguished colors. As a result, people who suffer from dichromacy may have trouble processing the information. People who are missing L cones and M cones cannot distinguish red from green on the color scale. People who lack S cones also find trouble distinguishing blue from green and yellow from violet. In addition, if the range of color scale is too diverse, normal people also find it hard to categorize and process the information.

In the Color Scale B, we see that it contains different tones of gray (white is an extremely light gray and black is an extremely dark gray). This grayscale is not much distinguished, apart from the saturation. Thus, it would cause some delay when our eye processes the visual information and interpret those statistics.

Exercise 4

Design a glyph that enables the pre-attentive perception of at least three variables (discrete or continuous). How many variables can you represent with your glyph, and what can you say about how the different variables are perceived?

Answer

Figure 10 below indicates the design of a glyph that enables the pre-attentive perception of at least three variables. The design was created using PowerPoint with a simple symbol.

Figure 8. Simple Glyph Design

The glyph design aligns with three variables which are color, size, and orientation. The number of variable combinations that can be represented by this glyph is calculated as

$$Combinations = Number\ of\ colors \times Orientation\ Steps \times Size\ Steps$$

When it comes to pre-attentive processing, we typically have 4-8 resolvable steps in each dimension. In our case, if we assume that the number of resolvable size and orientation steps are four (Ware, Collin, Information Visualization). Meanwhile, we use a total of three different colors used to represent the glyph. When plugging the numbers inside the equation, we have

$$Combinations = 3 \times 4 \times 4 = 48$$

Since size and color are distinguished features, our eyes perceive the color and size independently from each other. Meanwhile, the size and orientation are more integral as these are perceived together during visual processing.