Student Name : Nguyen Xuan Binh
Student Number: 887799

Information Visualization Assignment 2

Exercise 1 (10 points)

Take a look at the EU Open Data Portal database at https://data.europa.eu/euodp/en/data/group.
From there, download one of the open data sets and visualize some interesting phenomena in the data. 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 printer. The use of available visualisation tools—e.g., R, Python (with libraries like Matplotlib), Matlab, Illustrator, (Open) Office, etc.—is encouraged. However, implementing your own scripts and small programs for processing the data will be necessary

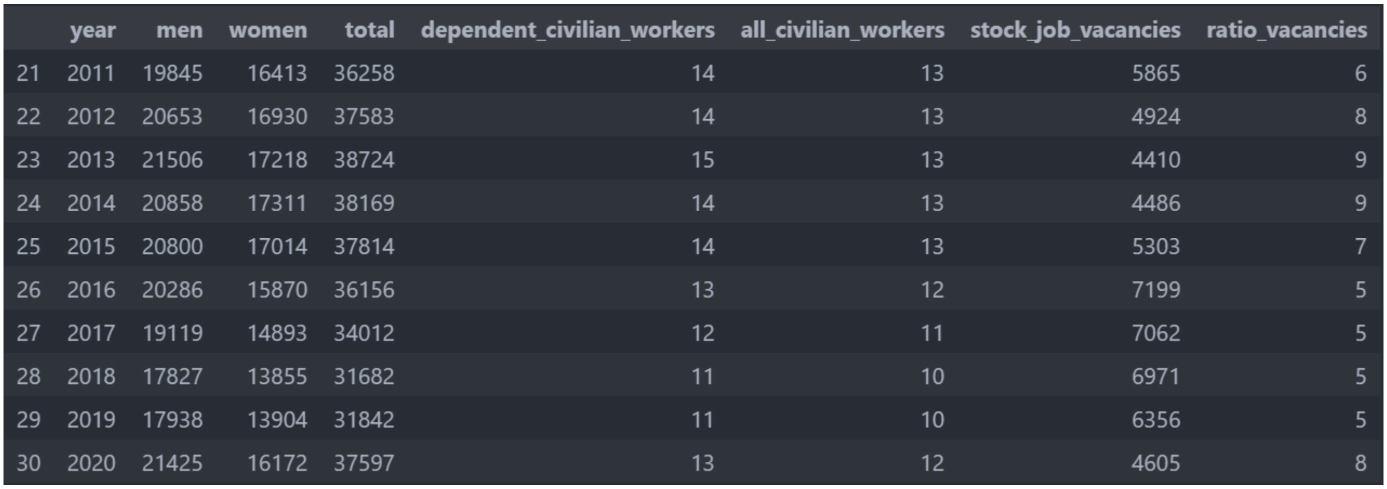
- Data topic and source

The dataset I chose in this exercise is the unemployed, unemployment rate and vacancies since 1990 in Germany from the EU open data portal database. It is collected and published by the Dortmund publisher and its latest update is the year 2020.

- Overview of the datasets:

The dataset has 8 columns and 31 rows in total. This is a relatively small dataset that is easy to handle for information visualization tasks without much data preprocessing. This dataset's columns are in German, which needs to be translated into English manually.
- year (jahr): The year of unemployment records in Germany
- men (männlich): Number of jobless men by year
- women (weiblich): Number of jobless women by year
- total (insgesamt): Total number of jobless people
- dependent_civiianl_workers: Unemployment rate based on dependent civilian workers (Arbeitslosenquote bezogen auf abhangige zivile Erwerbspersonen)
- all_civilian_workers: Unemployment rate based on all civilian workers (Arbeitslosenquote bezogen auf alle zivilen Erwerbspersonen)
- stock_job_vacancies: Stock of reported job vacancies at the end of the month (Bestand gemeldeter offener Stellen am Monatsende)
- ratio_vacancies: Ratio of unemployed/registered vacancies (Verhaltnis Arbeitslose/gemeldete offene Stellen)

The overview of dataset is shown below for the latest decade



	year	men	women	total	dependent_civilian_workers	all_civilian_workers	stock_job_vacancies	ratio_vacancies
21	2011	19845	16413	36258	14	13	5865	6
22	2012	20653	16930	37583	14	13	4924	8
23	2013	21506	17218	38724	15	13	4410	9
24	2014	20858	17311	38169	14	13	4486	9
25	2015	20800	17014	37814	14	13	5303	7
26	2016	20286	15870	36156	13	12	7199	5
27	2017	19119	14893	34012	12	11	7062	5
28	2018	17827	13855	31682	11	10	6971	5
29	2019	17938	13904	31842	11	10	6356	5
30	2020	21425	16172	37597	13	12	4605	8

The purpose of this exercise is to achieve multiple cornerstones of good visualization practices for this dataset, which are the Tufte's principles, Gestalt's laws, pre-attentive processing and appropriate colors for the human's eyes.

Recap:

- Tufte’s principles state that five aspects improve the visualization's quality. The principles are

- maximizing the data-ink ratio
- reducing the chartjunk
- multifunctioning graphical elements
- adequate data density and using small multiples
- ensuring the aesthetics and techniques

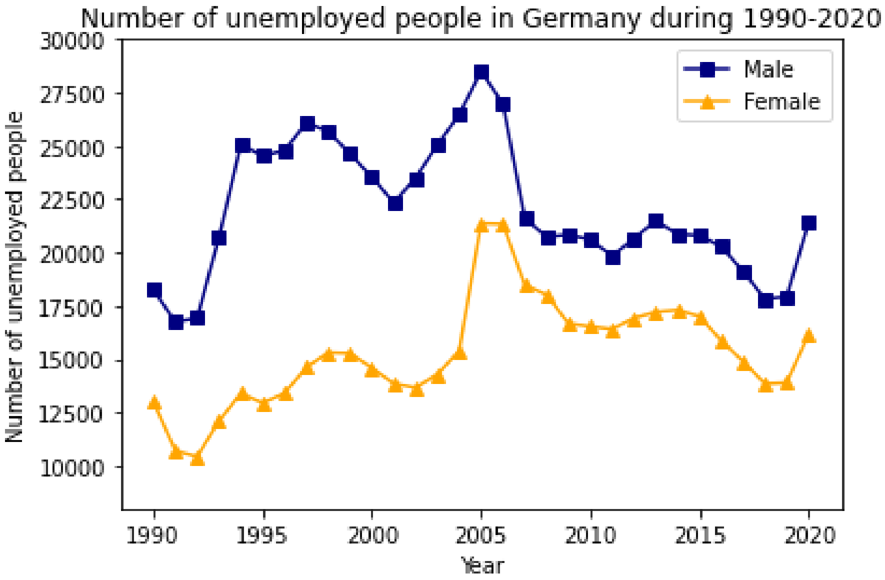
- Gestalt's laws translate directly into design principles of visual displays. The laws are:

- Similarity
- Good continuation
- Proximity
- Symmetry
- Closure
- Relative size
- Common fate
- Patterns from motion
- Animation and perception of shapes
- Causality

- Pre-attentive features are some visual objects are processed preattentively (popped out) before the conscious attention. The pre-attentive processing speed is independent of the number of distractors. Pre-attentiveness are highly desirable to catch the viewer's attentions.

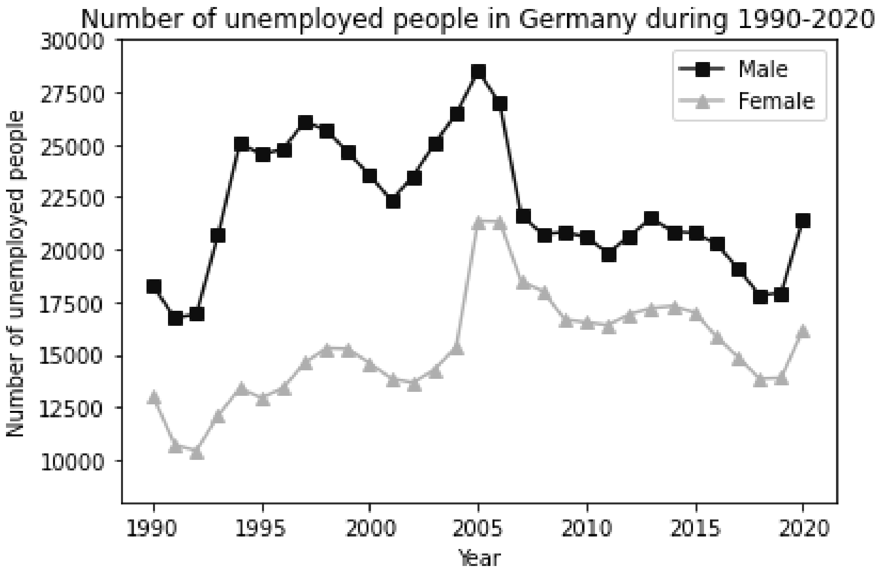
- Colors theory for human eyes, which governs the color schemes such that they avoids:

- Inability to distinguish colors by color blind people
- Simultaneous brightness-contrast effect
- Unsuitable color scale and absolute lightness perception
- Undistinguishable colors when the figures are printed in grayscale.

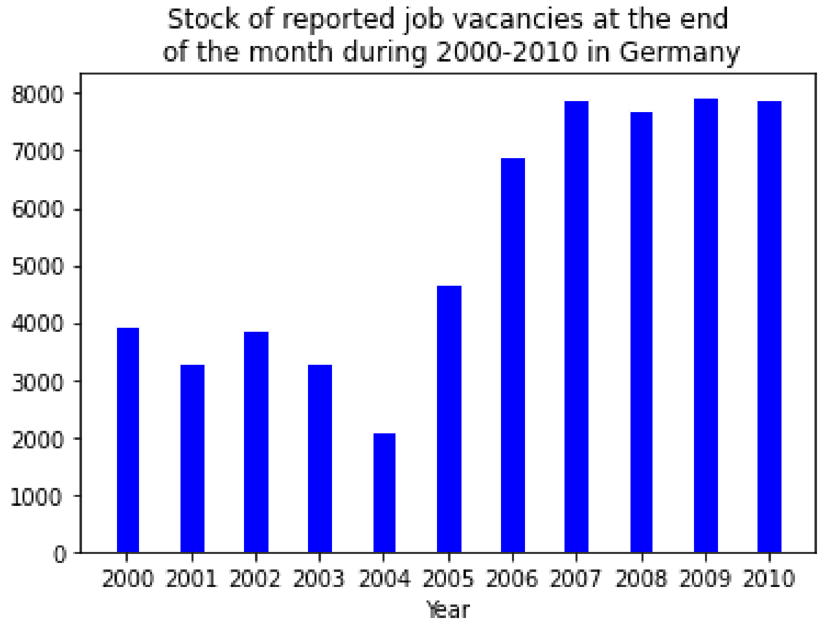


First is the graph for the number of unemployed men and women during the three decades in German. This line graph obeys many Tufte's principles, such as maximizing the data-ink ratio (the total number of unemployed people is not plotted, no grid lines, no redundant data). There is no chartjunk and the time series lines are plotted against each other in close proximity for easy comparison. The graph has a simple design and has the golden ratio.

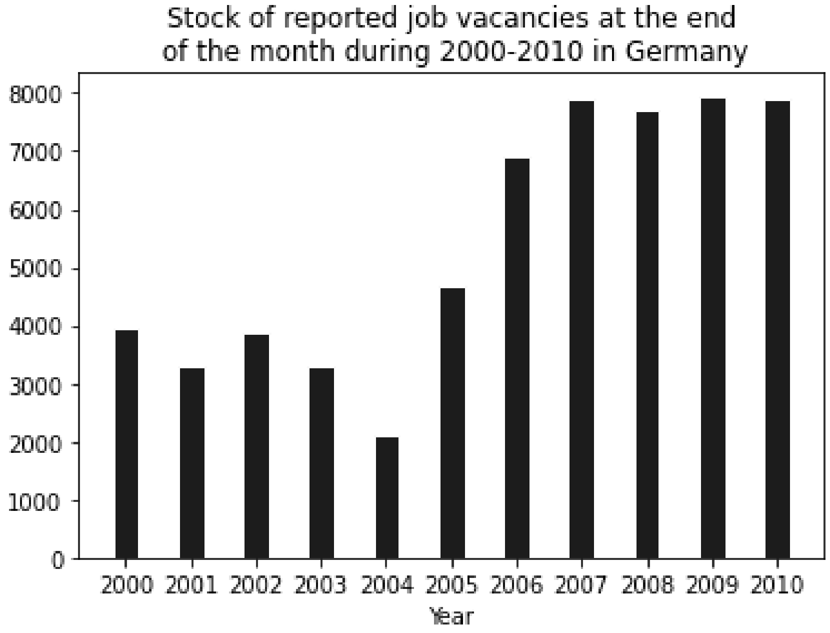
This line graph also has the blue-yellow channel, which is effective for color blind people that cannot distinguish red and green colors. Yellow and blue are the natural contrasting colors that human eyes can preattentively distinguish. This graph also obeys the Gestalt's law of continuity



When plotted in grayscale, the brightness of the two lines also differ to avoid confusion. The preattentiveness is further enhanced with the different markers on each line graph in the case that the printed lines are not contrasting enough. As a result, this line graph can caught the viewers' attentions preattentively with and without colors.

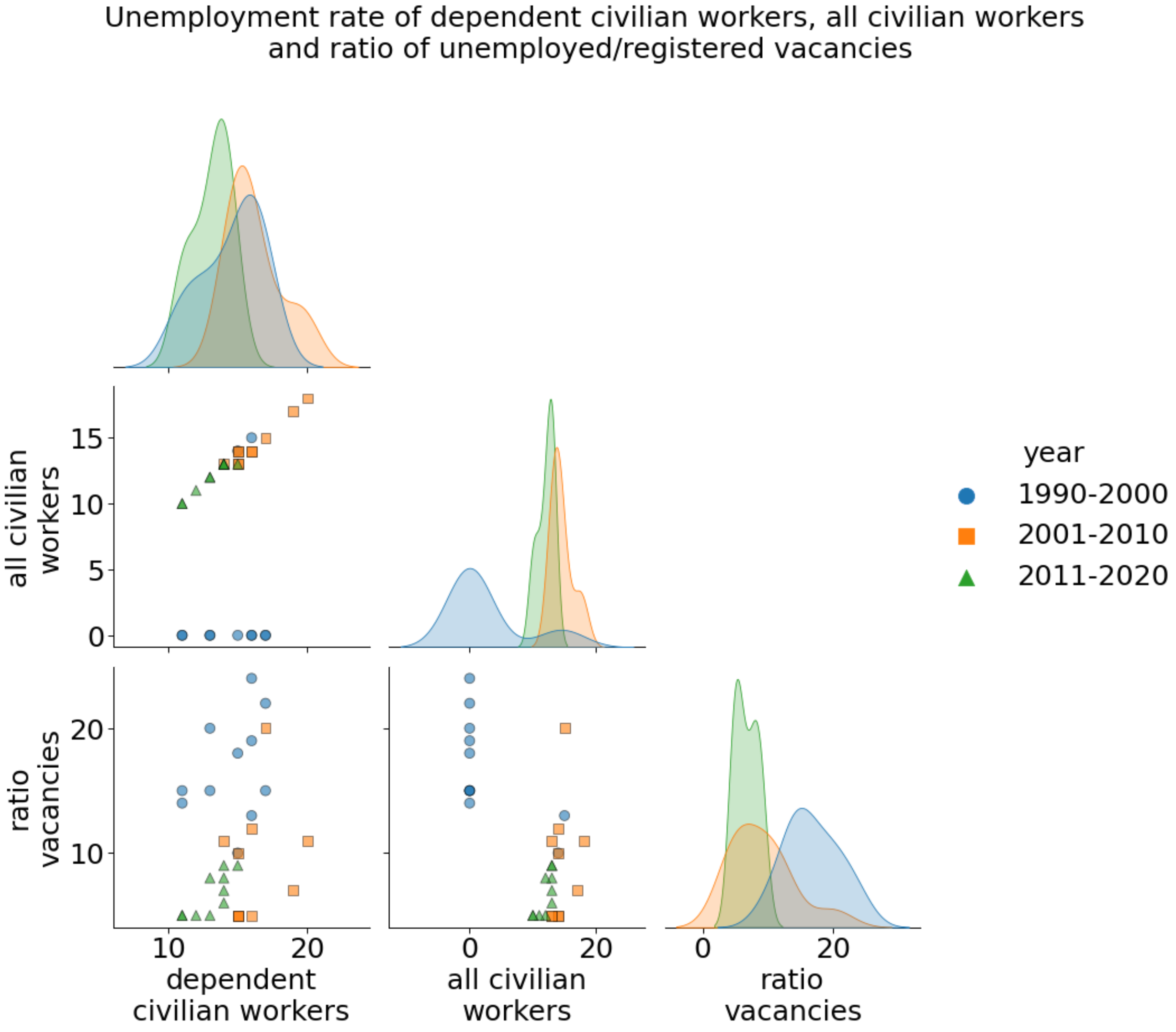


The second graph is about the job vacancies stock values in one decade. This is a simple bar graph that also obeys the Tufte's principles, which are no chartjunk, maximum data-ink ratio and has an ideal proportion. This graph also features the Gestalt's laws of symmetry, which is that all bars are evenly spaced. The bars are symmetric along the horizontal axis, which makes the pattern comparisons easier. Furthermore, there is only color in this graph (blue), so there is no problem of simultaneous brightness contrast effect nor being confused by color-blind people,



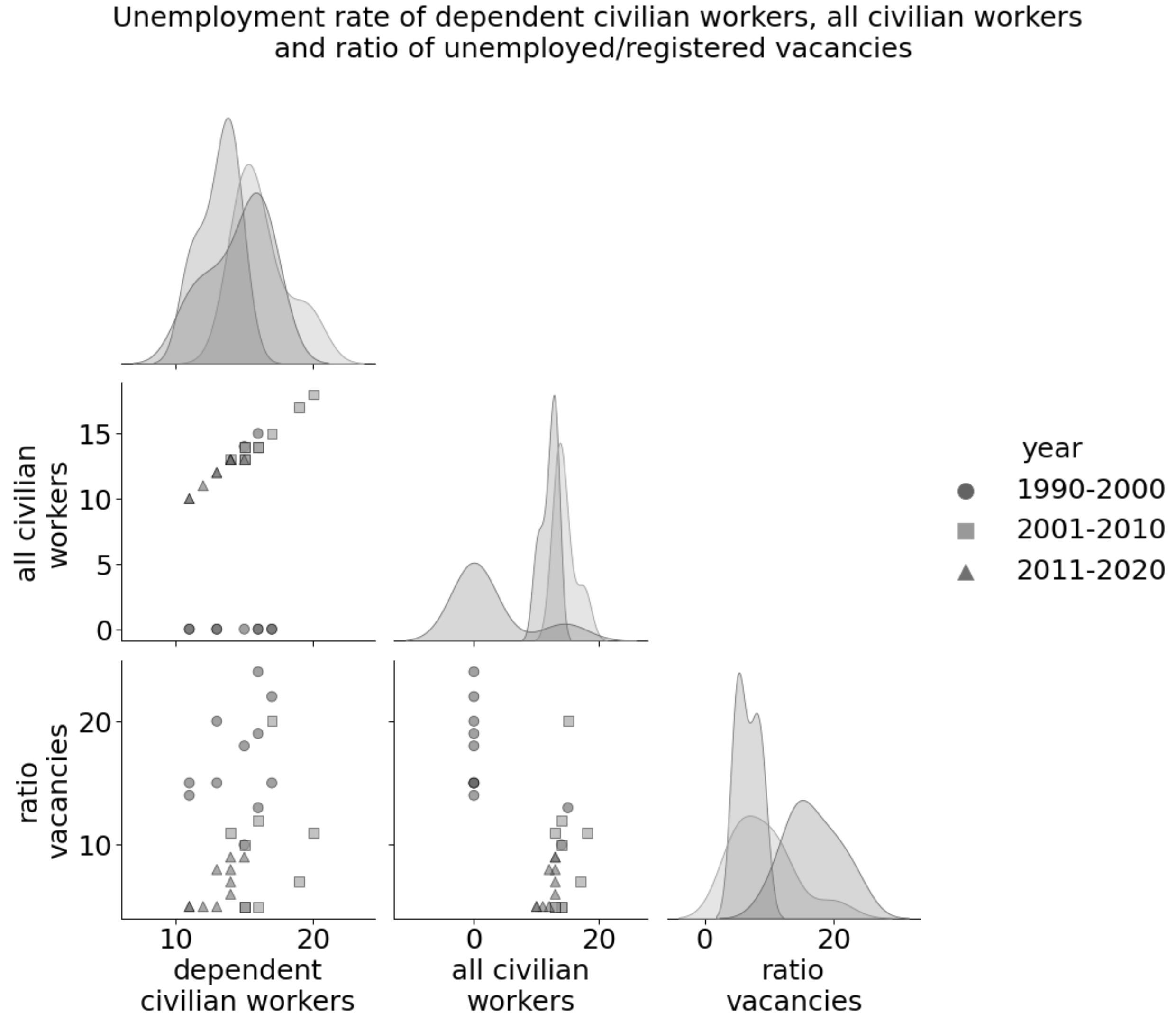
When plotted in grayscale, the graph can still be easily perceived like the colored version, so both of the colored and grayscale versions are easy to read and perceived. .

Finally, the trellis graph is plotted for three features, the ratio of dependent civilian workers, of all civilian workers and the unemployed/registered vacancies.



This graph obeys the Tufte's principles of maximum data ink ratio (only the lower half triangle is plotted instead of the whole graph), no chartjunk and enabling the small multiples comparison between the features. The scatterplot features the markers with both varying colors and shape, which allows pre-attentive preprocessing. This is because each pair of color and shape datapoint corresponds to one decade, so this is not a conjunction search but rather a single combination of dimensions. For example, if all of the markers are plotted with a circle shape, then lots of circles will be overlapping on the scatterplot and it will be hard to figure out which decade the datapoint belongs to if their hues are nearly similar in. This trellis plot also obeys the Gestalt's laws of proximity, where datapoints of a decade tend to cluster together, enhancing the separation between the decades and the emphasizing the connections between the years.
Finally, the color channel of the trellis is only blue/yellow channel with additional green color. Therefore, reading the graph is also possible for color blind people without L/M cones.

Finally, the trellis graph plotted in grayscale is also easy to be processed. Here are the justifications for this statement.



This is where the role of shape of the scatterplot points shine. When printed in greyscale, if all of the datapoints have the same shape, it is extremely challenging to figure out which decade each datapoint belongs to. However, since the decade markers are also distinguished by shape, the grayscale trellis enables pre-attentive preprocessing by the shapes when the colors become ineffective. The shape dimension also cater to color-blind people if they have a hard time deciphering the relative luminance of the grayscale. The diagonal of KDE, even though plotted in greyscale, can also be perceived as a continuous distribution thanks to the continuity in Gestalt's law, which states that human visual system tends to perceive lines continuously instead of abrupt changes between the two irregular shapes. In conclusion, both the colored and grayscale trellis obey the main visualization principles.

Exercise 2 (2 points)

Discuss the simultaneous brightness and contrast eect shown in Fig. 1 and explain it using the Difference of Gaussians model. Your answer does not have to contain any mathematical calculations. Instead, try to give a clear explanation why the rectangles appear to be of a different gray tone while they in fact are all of the same “color.” What are the implications of this phenomena for the design of color scales?

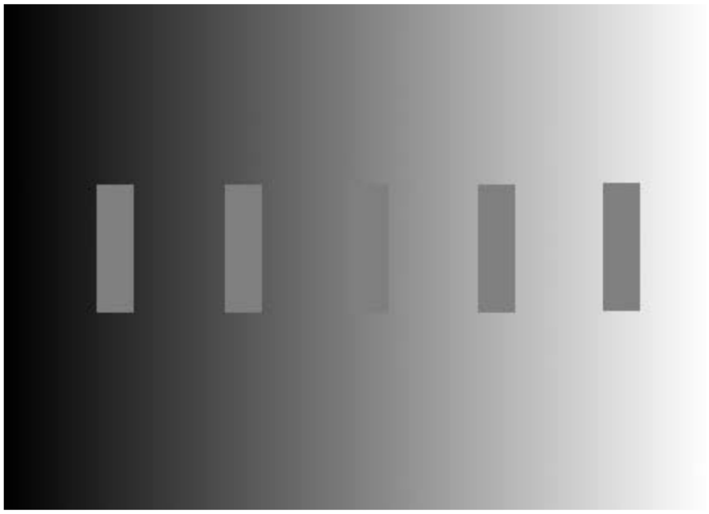
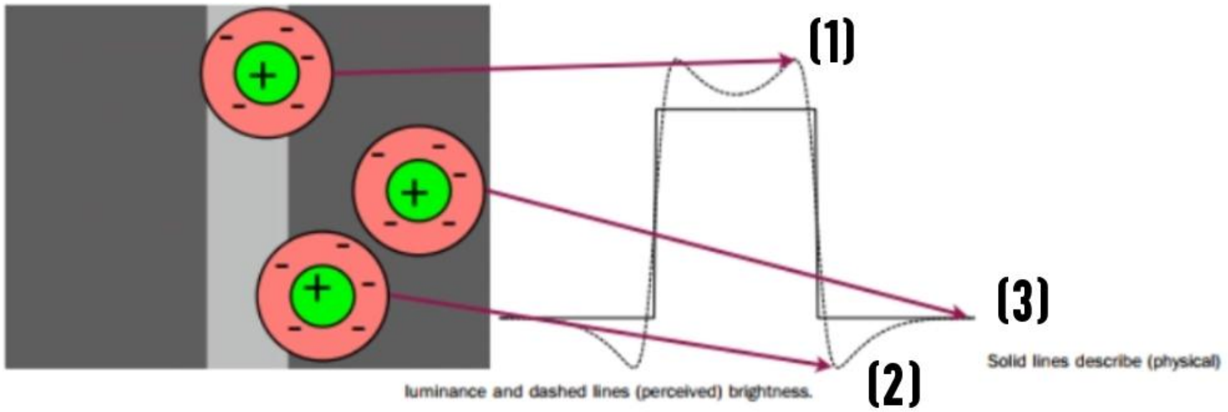


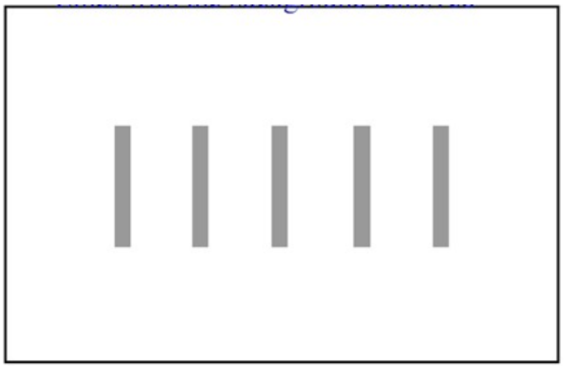
Figure 1. Simultaneous brightness and contrast effect.

Figure 1 features five bars evenly distributed on the grayscale background of increasing brightness from left to right. According to the claim, the five bars share the same brightness, but they appear to have different shades to the human's eye. From my point of view, the first bar in the left appears to be brightest and the last bar on the right is the darkest. This illusion can be explained by the Difference of Gaussians (DoG) in that the human eyes' receptive fields behave like DoG as retinal ganglion cells are organized with circular receptive fields. When light falls at the center of the receptive field, it emits pulses at increased rate, which is called excitation. When light falls off the center of the receptive field, it emits pulses at a lower rate, which is lateral inhibition. Mathematically, DoG is a linear filtering model of the receptive field in different layers of the visual pathway, which mimics how neural processing in the retina of the eye extracts details from images destined for transmission to the brain (source: Kenneth R. Spring - Scientific Consultant). This figure below shows how physical and perceived brightness differs in an example of Chevreul illusion


(1)
(3)
Solid lines describe (physical)
luminance and dashed lines (perceived) brightness.
(2)

In the simultaneous brightness-contrast effect image, the first bar is placed against the darkest background, which creates lots of visual excitation as the main source of light is the bar that falls at the center of the receptive fields. This causes the first bar to appear brighter than its true shade, which corresponds to position (1) on the line graph. On the other hand, the last bar is placed against the brightest background, which induces lateral inhibition as the main source of light is the background that falls at the edge of the receptive fields. This causes the last bar to appear darker than its true shade, which corresponds to position (2) on the line graph. Therefore, the DOG model can explain the difference between physical luminance and perceived brightness in the simultaneous brightness-contrast effect

This is the version without the contrast brightness. This time the viewers have no trouble believing that all the five bars share the same shade.



The implications of this phenomena for the design of color scales: when designing a numeric scale for comparison using the grayscale, it is best that all of the image contains only contains the information of numeric values. Any shades carrying other meanings such as nominal or labelling use must be avoided. This is because when grayscale is chosen as a nominal data, the shade of the objects will be perceived differently if the background has varying brightness. For best practice, we can just avoid using grayscale for nominal/labelling tasks.

Exercise 3 (1 point)

Fig. 2 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.

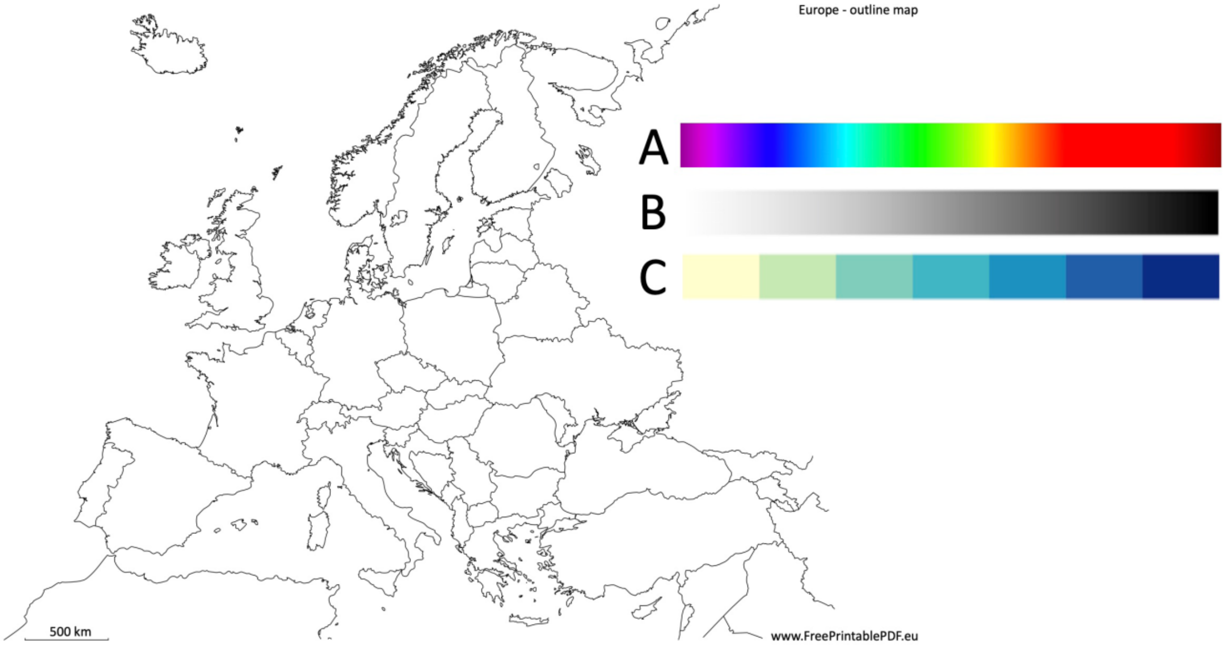


Figure 2. Three colour scales (A: Spectral, B: Greys, C: YlGnBu) and map of Europe with country borders.

Regarding color-blindness, males have 1 X chromosome and 1 Y chromosome, and females have 2 X chromosomes. The genes that can give red-green color blindness are passed down on the X chromosome. Since it's passed down on the X chromosome, red-green color blindness is more common in men. Particularly, the lack of L cones and M cones occur in 1-1.5% in men and 0.01% in women, which results in the inability to distinguish red from green. Lack of S cones (cannot distinguish blue/green and yellow/violet) is also existing but in very small percentage of the population. Therefore, we should not choose a color scale that has both red and green hues to avoid the problem of color-blindness.

Regarding contrast sensitivity, physical luminance and perceived brightness can be quite different for human vision, as shown previously in Exercise 3. Therefore, the color scale should not have the simultaneous brightness contrast effect. Additionally, people tend to divide colors to a few basic categories and the closer the color is to the "pure color", the easier it is to remember it. This is the opponent color theory, where 6 elementary colors are arranged perceptually, where cone signals are transformed into 3 distinct channels: black-white (luminance), red-green, and yellow-blue. Therefore, the color scale for comparison should be discrete (categorical) and belongs to only one channel.

Regarding lightness perception, human visual system is adapted to illumination levels of six orders of magnitude. The absolute illumination levels are essentially ignored. The lightness perception is extremely relative due to adaptation and lateral inhibition. Therefore, we should choose a scale for numeric comparison that does not have too many hues on it as the absolute luminance is irrelevant to the human eyes.

Given these observations, the best color scale for statistical data for comparing the countries is the color scale YlGnBu (C). Here are the justifications:

Spectral scale A: This color scale is unfit for comparison simply because it has too many hues and it will be very hard for human eyes to process the relative difference because the colors do not belong to any channel. Additionally, it has both red and green color, which can cause a problem for color-blind people whose eyes lack L and M cones.

Gray scale B: This color scale is unfit for comparison because the order of luminance magnitude is not clearly bounded, but rather a continuous range of brightness. This makes it very challenging to make comparisons as human eyes need to process the relative luminance of the grayscale to actually infer the data. Additionally, this scale may also suffer from the simultaeous contrast effect if the background has anything other than pure white or pure black.

YlGnBu scale C: This color scale consists of yellow and blue colors with varying brightness. This scale meets all of the requirements listed above. It does not have both red and green, which avoids the problem of color-blindness. Since the color scale has completely different colors from the background (black/white), the simultaneous contrast effect is not existing. Furthermore, this color scale belongs to only 1 channel, which is the yellow-blue channel and consists of discrete hues. Finally, this scale only has a few colors, so it is fit for statistical data relative comparison.

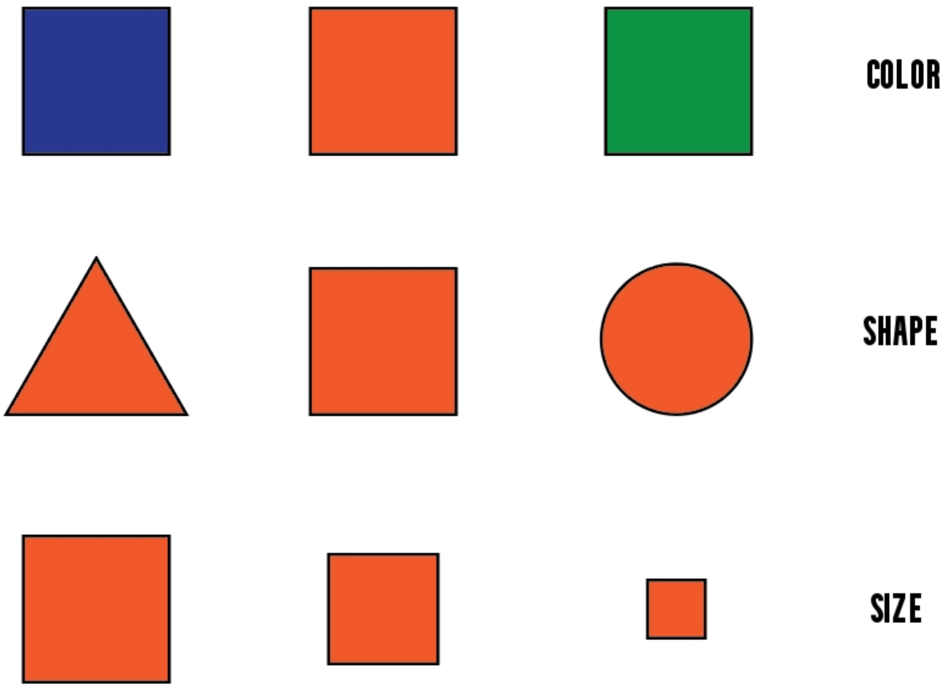
Exercise 4 (2 points)

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?

By definition, glyphs are symbols used to represent multivariate data, where a single glyph corresponds to one sample in a data set. The point of glyphs is to be perceived pre-attentively by viewers with multiple features of the glyph.
Some rules of thumb rules for designing a glyph that enables pre-attentive processing are:

- Data variables should (usually) be mapped to pre-attentive features
- Try to use separable channels in stead of integral channels. Separable features are perceived independently of each other (e.g., size and color), whereas integral features are perceived holistically (e.g., a width and height). Integral visual dimensions interfere with each other, so we sHould use separable dimensions instead.
- have 4-8 resolvable steps in each dimension (e.g., the number of size steps we can easily distinguish is ~4).
- Conjunction searches are usually not pre-attentive
- One can effectively display only limited number of visual variables, with limited accuracy

The figure below is my simple glyph design that encodes data for three discrete variables. The image was created with Adobe Illustrator.



Supposed that a dataset has 3 discrete features, where each sample is assigned to an object that can be located in the Cartesian spatial coordinates (X, Y, Z), then the glyphs above can represent at most (3 colors) x (3 shapes) x (3 sizes) = 27 different combinations for all samples. In fact, three is the recommended dimensionality for color, shape and size dimensions, as humans can easily distinguish up to around 4 resolvable steps. All of these features are separate dimensions that enables easy pre-attentive processing by viewers, because colors, shapes and sizes do not influence the value perception of each other, unlike length/width or motion/orientation. Therefore, this glyph design also does not have the problem of integral visual dimensions. Furthermore, there are no conjunction searches, as each dimension corresponds to one feature as proposed. As a result, for the sake of pre-attentiveness, there is only limited number of visual variables, which is 27.