WWU MÜNSTER

University of Münster
School of Business & Economics

Leon Cena

Data visualisation in the context of the Corona pandemic

Seminar Thesis

in the context of the seminar “Data Visualisation”

at the University of Liechtenstein

Principal Supervisor: Ass.-Prof. Dr. Johannes Schneider
Associate Supervisor: Ass.-Prof. Dr. Johannes Schneider

Presented by: Leon Cena [FS 210 128]
+49 176 24715761
leon.cena@uni.li

Submission: 23rd January 2022

Contents

1 Introduction

Hardly a day goes by these days without a reminder that we are in a pandemic. We restrict our freedom in order not to endanger our fellow human beings. Many everyday things are therefore no longer possible. New everyday activities have now become established. Regularly one looks in every possible source into the pandemic course. One would like to be informed, no one would like to endanger the own person or the fellow persons. Many people probably could not have imagined hearing about 7-day incidences on a daily basis before the pandemic.

Statistics, metrics and visualizations have managed to become a central aspect of life. Measures, travel restrictions and one's own behaviour are significantly influenced by these. However, the situation cannot be explained with such certainty. Data at Corona times is strictly accompanied by data noise. One works with statistics, which are known to be error-prone, and makes decisions based on them. At the same time, not everyone has a statistics education behind them. Many people did not have these skills and are now confronted with them every day. Therefore, rumours quickly arise that are based on misunderstood statistics. Especially in times of Fake News it becomes clear that it is indispensable to have a basic knowledge of statistics if you have to deal with the current topic of the pandemic.

The subject of this work is not solely the creation of diagrams. Rather, it is about the complete process behind each data visualization, but additionally applied to the current pandemic situation.

To start the work, the relevance and problems of data visualization are answered. Thereby the questions about the necessity and the surprising interest of normal people for statistics are answered. Special attention is given to the current problems, especially with regard to the aspect of uncertainty. Subsequently, the theoretical foundations of data visualization are laid. Helpful theoretical concepts are introduced which can be used as a basis for the practical part. The main part of the thesis deals with the process of visualization. First, the data basis of this project is presented and outlined and then the data preparation is discussed. Afterwards different visualizations are created and discussed. There is no strict coherent guiding question but rather the aim is to show as many facets as possible with the help of the database. The statistics of the number of cases, years of death and vaccination figures play a central role. The context of the analysis is international. Both extreme cases and universal worldwide facts are visualized.

Finally, the created visualizations are discussed, partly with the help of concepts from the literature.

2 Relevance and problem of visualizing the Corona pandemic

The Corona pandemic has accompanied mankind for quite some time. Since then, numerous people have been dealing with statistics more intensively for the first time. Timcke and Schneider (2021), German journalists, report that readers of their newspaper are extremely interested in learning more details about the statistics behind the pandemic including the calculation methods of the key indicators. People actively question the figures and ask specific questions, this was previously done to a large extent by professional statisticians. People actively strive to achieve higher data literacy. According to Frank et al., data literacy is the ability of non-experts to make logical use of significant data and, consequently, to understand it correctly (cf. Frank et al. 2016, p. 5).

This phenomenon is understandable, as there has not been such a relevant and data-driven day-to-day issue for quite a long time. This can be explained by the fact that Corona statistics have a direct influence on people's everyday lives. First, personal well-being is affected as people worry and worry about the pandemic. On the other hand, key figures are often used by the government to determine measures that have a direct influence on people's everyday lives (cf. Timcke and Schneider 2021, p. 102). There are also effects on people's actions. Adherence to preventive measures aimed at controlling the pandemic results from correct and comprehensible risk communication by the government and the media (cf. Loss et al. 2021, p. 294). Visualisations are a way of achieving that goal since they transmit key indicators which are primarily the basis of the risk measurements (cf. Timcke and Schneider 2021, p. 102).

The omnipresent uncertainty of the Corona statistics must also be taken into account. They should therefore be treated with caution. The data does not necessarily reach the data processors in a well-maintained form. Manual adjustments are often necessary. The definitions and regulations on which statistics are based can also change. Visualisations based on current measures, such as the threshold for lockdown, lose their validity as soon as the regulations behind them change. Central definitions are not necessarily equivalent across countries. For example, in Italy, Corona infected people who remain in quarantine at home were not counted in the statistics for a limited period of time. This results that countries are not directly comparable in terms of Corona statistics. However, cross-country visualisations exist as a way to give people a rough overview. A better option is not necessarily available (cf. Timcke and Schneider 2021, p. 106). It is therefore essential for good risk communication that such uncertainties of scientific results are addressed transparently (cf. Loss et al. 2021, p. 296).

Different fundamental characteristics of different countries also contribute to the fact that they cannot be easily compared or visualised together. The data of the different countries have different subtleties. In some countries, data is recorded in extremely fine detail, such as the varying statistics of the German federal states. Others, on the other hand, often only show detailed statistics for the entire country. Characteristics such as the area of a country play a more important role depending on the information. For example, Liechtenstein would be much harder to notice on a world map visualisation than Russia. The same pandemic situation could therefore have a completely different impression in a different country. In that case, a particularly large prominence on the world map could make the situation of Russia appear particularly threatening (cf. Timcke and Schneider 2021, p. 108). To combat these problems, the German Federal Statistical Office, together with Eurostat, has developed a European dashboard to make uniform, up-to-date and comparable data available to the public (cf. Schliffka 2021, p. 31).

In the course of the pandemic, various indicators were used to describe the situation. At the beginning of the pandemic, the number of recorded infections was presented more prominently, as it initially provided a good overview of the start of the pandemic. However, this figure alone cannot accurately represent how the situation is improving in concrete terms of a time-series. After all, the number of infections is continuously increasing. This would give a negatively distorted picture of the situation, as one has no information about recovered cases. Consequently, these ratios were eventually used as a pair. In the meantime, incidence has become established as a key performance indicator (cf. Timcke and Schneider 2021, p. 108).

As the pandemic progresses, the potential for possible misunderstandings of visualisations also increases. Not every person reading a data visualisation knows and understands the meaning of metrics, such as incidence or R-number. Readers might misinterpret different values in different sources, justified by divergent metric calculation methods, as errors or even deliberate lies. This again reinforces the importance of data literacy. Some other readers might not understand that numbers are not always exact. However, this is often due to technical reasons (such as the quality of the data source) or organizational reasons (such as administrative reporting delays) (cf. Timcke and Schneider 2021, pp. 109-110). For these reasons, structural problems may arise in data collection. If no new data is recorded on weekends and holidays, this could distort the data for the following working day. In such cases, the audience needs to take a particularly good look at the resulting visualisation in order to interpret it correctly. However, due to the everyday relevance of the pandemic, this willingness is present in the audience (cf. Timcke and Schneider 2021, p. 110, 112).

The increasing relevance of the pandemic is accompanied by a flood of information from various media and reports. It is understandable that the topic also resonates in social media and generates discussions. Not everyone succeeds in filtering correct and relevant information from the large quantity and thus speculations, rumours and false reports are also spread. Therefore, it can be said that we are also in an "infodemic" (cf. Loss et al. 2021, p. 298).

3 Theoretical foundations of data visualisation

3.1 Perception and data visualisation

Before this work is dedicated to the technical part, we examine the theoretical foundations of visualisation in more depth. First we look at the perception and general aspects of data visualisation. Afterwards we deal with the guiding rules of the "Grammar of Graphics" and the approach of data analysis.

First we deal with the concept of perception. Perception is defined as the way of receiving information through any kind of environment that might be limited to the mind as well. (cf. Graham 1869, p. 1). For example, I see a car, the car is presented to my mind and then finally perceived. When I hear the sound of car I perceive it as well. An analysis of the visual perception of data visualisation can help why some plots can convey the intended message correctly and some fail that objective (cf. Healy 2018, chapter 1.3).

Edges, contrasts and colours play an important role in terms of perception. When forms share a boundary we automatically overestimate the contrast of those forms. Humans by instinct observe the surrounding of objects as well. This results in a rather relative than absolute assessment of features. In the context of two geometric forms one bases their comparison on the relative difference in terms of the colours and does neglect the absolute values. The checkershadow illusion intensifies on this effect. Even though one might know the result of certain experiment it is not possible to trick the own perception since the process is executed unconscious. One can apply this concept on colours as well. Human recognize contrasts better when there is only one colour, with different shades, involved. These effects underline how easy it is to be misled by a visualisation without even recognizing it and shows how important it is as a creator of visualisations to note the theory of perception while creating them (cf. Healy 2018, chapter 1.3.1).

The choice of colours is even more important. In addition to the brightness one can classify a colour into two component, the hue and the chroma. The first one denotes prevalent recognized colour (for example dark red and light are both recognized as red). The latter one denotes the intensity of the given colour. When mapping numerical variables to colour it is trivial that we try to choose colours with the same colour-distance. Theoretically this can be done through tuning the RGB-components. Practically this is not necessarily realizable because the human perception is not uniform. The visible range of chrome is dependent on the chosen luminance. The audience could start seeing scales of a plot in a wrong way because of that effect. To counteract that effect one can select colour schemes with uniform perception (cf. Healy 2018, chapter 1.3.1).

Preattentive search deals with the facts that objects can occur more visible than other objects in an environment. That means that they stand out in a set of objects. This might even happen before one looks at it with consciousness. The degree of being in fact preattentive is subject of discussions in the context of literature. For the context of data visualisation, however, this existing standing out, which is also describes as "preattentive pop-out", is of primary relevance.

When trying to analyse that effect we can classify two channels of recognition, shape and colour. Its is easier to detect a different colour in a visualisation than a different shape thus the pop-out of the first one seems to be higher. Other possible channels are angle, position (in terms of a proper structure and differentiation) and opacity. When using a sample with high number of items multi-channel searches are even more difficult (cf. Healy 2018, chapter 1.3.2). We can observe these effects in Figure 1. When only using the channel of shape the search starts to become harder. Even more complexity is added when the colour is added to the shape thus demanding a multi-channel search. Readers might need to search through the whole picture to find the blue circle. These findings must be taken into account when creating data visualisations.

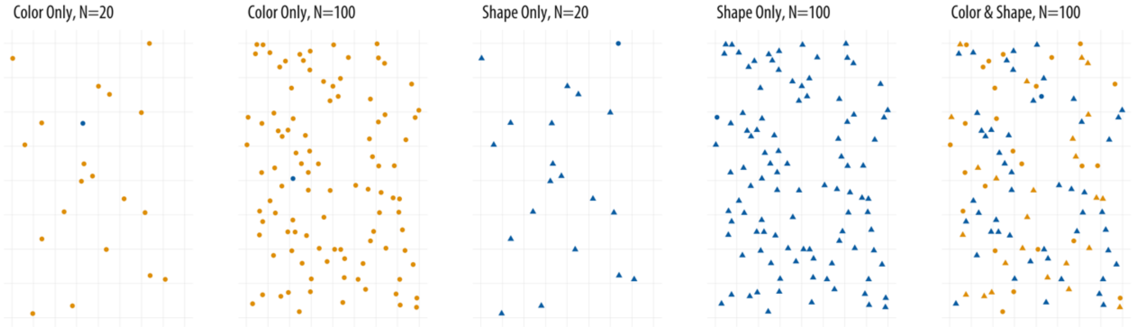


Figure 1 Example of the "preattentive pop-out": Searching for the blue circle becomes progresssively harder (Extracted from Healy 2018, chapter 1.3.2)

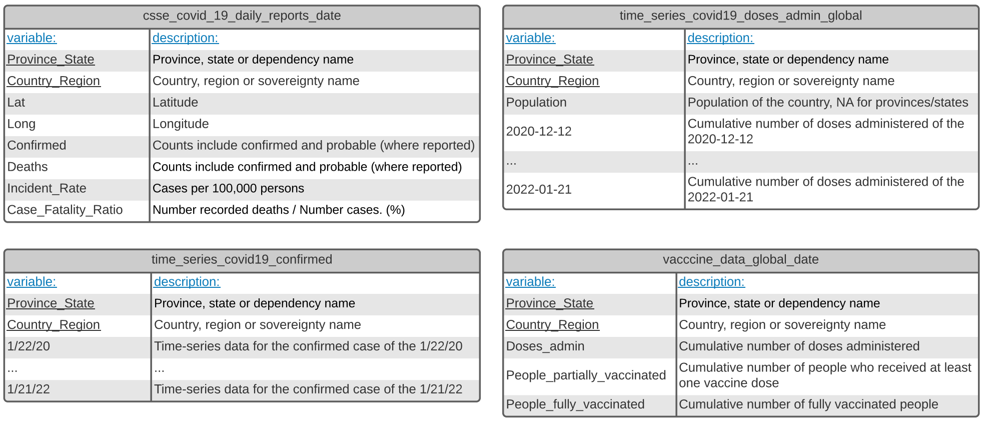
When looking at a visualisation people permanently try to identify structure. Even though there might not be much of information that is conveyed by the data one does draw predictions about relationship. The principles underlying that are called "gestalt rules". These rules are not fixed rather they specify how humans draw conclusions of group affiliation. Healy describes seven rules in this regard.

First, there is the rule of proximity. This describes that objects that have little spatial distance to each other appear to be connected. The same applies for objects that share a similar look. If you connect lenses visually, you also achieve the effect that they appear to be in relationship. For instance, this can be represented by a connecting edge. Continuity defines that to an extent hidden objects appear related. This rule goes hand in hand with the rule of Closure that defines that incomplete shapes can also be recognized as related. For example the reader might understand that a line can go under another line when they cross. Objects that match one or both of the last two rules could be for example recognized as being related with a structure that is in close distance thus it also involves the rule of Proximity. People can recognize whether an object is located in the background or foreground when there is a visible/perceived hierarchy. This is defined in the rule "Figure and ground". The last introduced rule is Common Fate. This describes that objects with the same direction are classified as related. According to that rule readers instantly understand charts of time-series in which objects are connected with a trend line and are read from left to right (according to Healy 2018, chapter 1.3.3). Knowing that this rules exist and are often relevant, even unconscious, helps to design charts while minimizing the risk of misunderstandings.

4 Visualisation of the Corona pandemic

4.1 Introducing the data source

In this paper, we will work with the extremely extensive data sets of the John Hopkinson University (JHU). For this purpose, the JHU aggregates various Corona statistics from the publications of the authorities of the respective countries. The JHU prepares this huge data collection as the first instance and in this way makes the JHU CSSE COVID-19 Data[1] available on GitHub on a daily basis. In addition, the JHU also provides a dataset containing statistics on the international vaccination data[2] of different countries. To clean the data we have to select interesting variables. Since we want to analyse the international statistics we do not pay attention to logistic variables like USA-unique IDs or unnecessary dates and logs. In the following diagram one can see the tables we will analyse with selected variables and their descriptions.



csse_covid_19_daily_reports_date

variable:	description:
Province_State	Province, state or dependency name
Country_Region	Country, region or sovereignty name
Lat	Latitude
Long	Longitude
Confirmed	Counts include confirmed and probable (where reported)
Deaths	Counts include confirmed and probable (where reported)
Incident_Rate	Cases per 100,000 persons
Case_Fatality_Ratio	Number recorded deaths / Number cases. (%)

time_series_covid19_doses_admin_global

variable:	description:
Province_State	Province, state or dependency name
Country_Region	Country, region or sovereignty name
Population	Population of the country, NA for provinces/states
2020-12-12	Cumulative number of doses administered of the 2020-12-12
...	...
2022-01-21	Cumulative number of doses administered of the 2022-01-21

time_series_covid19_confirmed

variable:	description:
Province_State	Province, state or dependency name
Country_Region	Country, region or sovereignty name
1/22/20	Time-series data for the confirmed case of the 1/22/20
...	...
1/21/22	Time-series data for the confirmed case of the 1/21/22

vacccine_data_global_date

variable:	description:
Province_State	Province, state or dependency name
Country_Region	Country, region or sovereignty name
Doses_admin	Cumulative number of doses administered
People_partially_vaccinated	Cumulative number of people who received at least one vaccine dose
People_fully_vaccinated	Cumulative number of fully vaccinated people

Figure 2 Entity Models of the data source with selected variables, key is underlined

4.2 Cleaning the data

After loading the CSV-Files we start cleaning the data. That means we choose the variables we want to keep. For the data sets we aggregate the numbers for every country. To do so we grouped the data set by the Country and applied the summarise function on each variable to sum the numbers up so that we receive the variables of the whole country. To reduce the size of the times-series data set we will reduce it down to the last 365 days.

The vaccination_global_data set offers particularly high granularity, as it also has the individual vaccination statistics of different states. However, we do not need this information

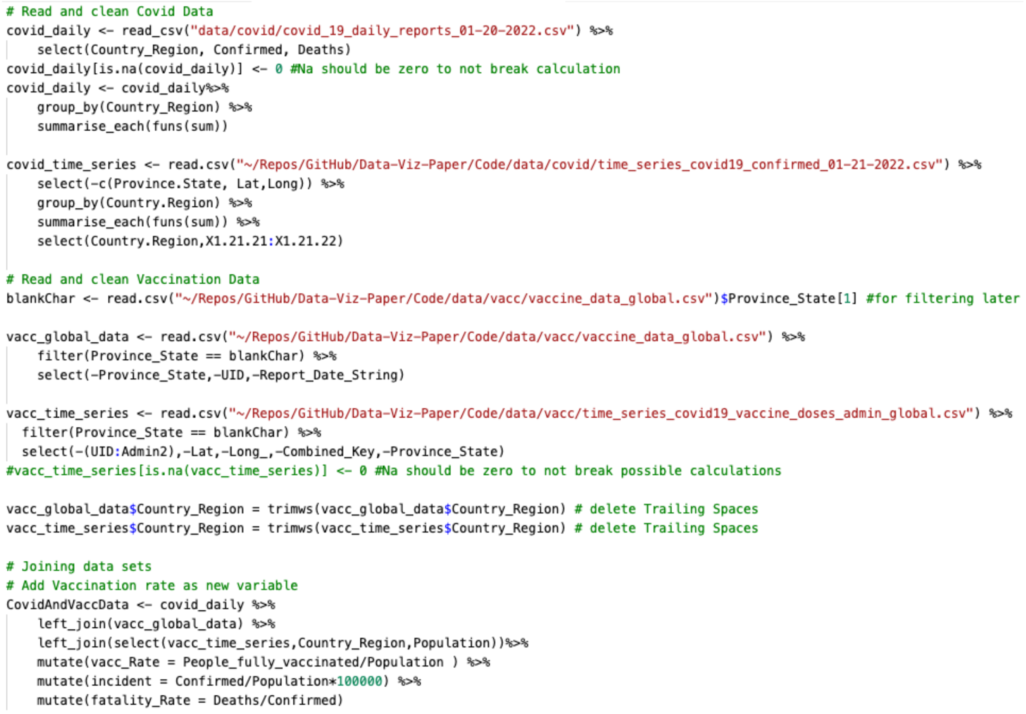
[1] https://github.com/CSSEGISandData/COVID-19

[2] https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/global_data

as we will visualise internationally. Therefore, the data will be aggregated again. This is easier this time because there is a data point for each country with the totals of the individual states. Therefore, we can filter out all data points from states so that only data points from countries remain. This can be applied to the vaccination-time-series data set as well.

In addition, we have created a large dataset by linking daily measurements of Corona infections with vaccination statistics. For this we used a leftjoin using the attribute of the country. To this we added the population of each country. This information resides in the vaccination_time_series dataset and is also added to the dataset via a leftjoin. With this data, we calculate the vaccination rate, incident rate and fatality rate for each country using the mutate function. When there is no population number the calculation of the vaccination rate and incident rate is not possible since the population is an element of the calculation of those figures.

Although cleaning and transforming the data set was quite strenuous, this has been indispensable. This is the only way we can reliably work with foreign data and create visualisations. This also helps to get familiar with the data set that has also positive effects when we need to calculate new statistics like the vaccination rate. Now we have exactly the data we want for (almost) every country. It becomes clear that data research and data preparation play a central role.



```r
# Read and clean Covid Data
covid_daily <- read_csv("data/covid/covid_19_daily_reports_01-20-2022.csv") %>%
    select(Country_Region, Confirmed, Deaths)
covid_daily[is.na(covid_daily)] <- 0 #Na should be zero to not break calculation
covid_daily <- covid_daily%>%
    group_by(Country_Region) %>%
    summarise_each(funs(sum))

covid_time_series <- read.csv("~/Repos/GitHub/Data-Viz-Paper/Code/data/covid/time_series_covid19_confirmed_01-21-2022.csv") %>%
    select(-c(Province.State, Lat,Long)) %>%
    group_by(Country.Region) %>%
    summarise_each(funs(sum)) %>%
    select(Country.Region,X1.21.21:X1.21.22)

# Read and clean Vaccination Data
blankChar <- read.csv("~/Repos/GitHub/Data-Viz-Paper/Code/data/vacc/vaccine_data_global.csv")$Province_State[1] #for filtering later

vacc_global_data <- read.csv("~/Repos/GitHub/Data-Viz-Paper/Code/data/vacc/vaccine_data_global.csv") %>%
    filter(Province_State == blankChar) %>%
    select(-Province_State,-UID,-Report_Date_String)

vacc_time_series <- read.csv("~/Repos/GitHub/Data-Viz-Paper/Code/data/vacc/time_series_covid19_vaccine_doses_admin_global.csv") %>%
  filter(Province_State == blankChar) %>%
  select(-(UID:Admin2),-Lat,-Long_,-Combined_Key,-Province_State)
#vacc_time_series[is.na(vacc_time_series)] <- 0 #Na should be zero to not break possible calculations

vacc_global_data$Country_Region = trimws(vacc_global_data$Country_Region) # delete Trailing Spaces
vacc_time_series$Country_Region = trimws(vacc_time_series$Country_Region) # delete Trailing Spaces

# Joining data sets
# Add Vaccination rate as new variable
CovidAndVaccData <- covid_daily %>%
    left_join(vacc_global_data) %>%
    left_join(select(vacc_time_series,Country_Region,Population))%>%
    mutate(vacc_Rate = People_fully_vaccinated/Population ) %>%
    mutate(incident = Confirmed/Population*100000) %>%
    mutate(fatality_Rate = Deaths/Confirmed)
```

Figure 3 Reading and cleaning of the data

4.3 Visualisation of the data

In the chapter on the relevance of the pandemic in the context of data visualisation, we came across the notion of data literacy and uncertainty. Both can now be demonstrated very quickly in the practical part. In the data, it often happens that entries do not fall. On the one hand, these exceptions have to be dealt with when preparing the data, as otherwise indicators could be calculated incorrectly. On the other hand, they must also be considered when visualising. A non-existent value can easily change the appearance of a visualisation and send the wrong message. That would be the direct example of the uncertain data situation.

The code powering the visualisations can be accessed on GitHub[3].

4.3.1 Countries with the most/least confirmed cases and their vaccination progress

The purpose of the first visualisation (cf. figure 4) is to show that graphs can be deceptive if they are not directly questioned. This visualisation is about 2 bar charts that are combined in one facet in order to compare them easily. In terms of data, the number of cases and the vaccination rate are visualised. In order to remain true to the theoretical principles, different shapes were not used and an easily recognisable colour palette was chosen, which offers an easily recognisable contrast. The purpose of this graph is to examine the indicator of detected Corona cases more closely and to relate it to another indicator, in this case the vaccination rate.

For this purpose, the 8 countries with the highest confirmed case rate were first output with R and included in the bar chart. If one wants to find the lowest case numbers, one encounters two problems. In some cases, no information is available from many small states. When there is no population number the calculation of the vaccination rate and incident rate is not possible since the population is an element of the calculation of those figures. Therefore, these are filtered out. Since we want to include the vaccination rate, data on this must also be available. Finally, the number of cases and the vaccination rate can be set to a minimum threshold in order to exclude outliers, such as island states that are not known. The vaccination rate is shown by the colouration of the rectangles.

Another possibility is to determine the vaccination progress not by the exact colour but by a category variable. An artificial variable was introduced that assigns a category to the individual data records for the purpose of vaccination progress. This has the advantage that one can use other types of visualisations that are based on categorical variables. The

[3] https://github.com/leoncena/Data-Visualisation-submission-Paper

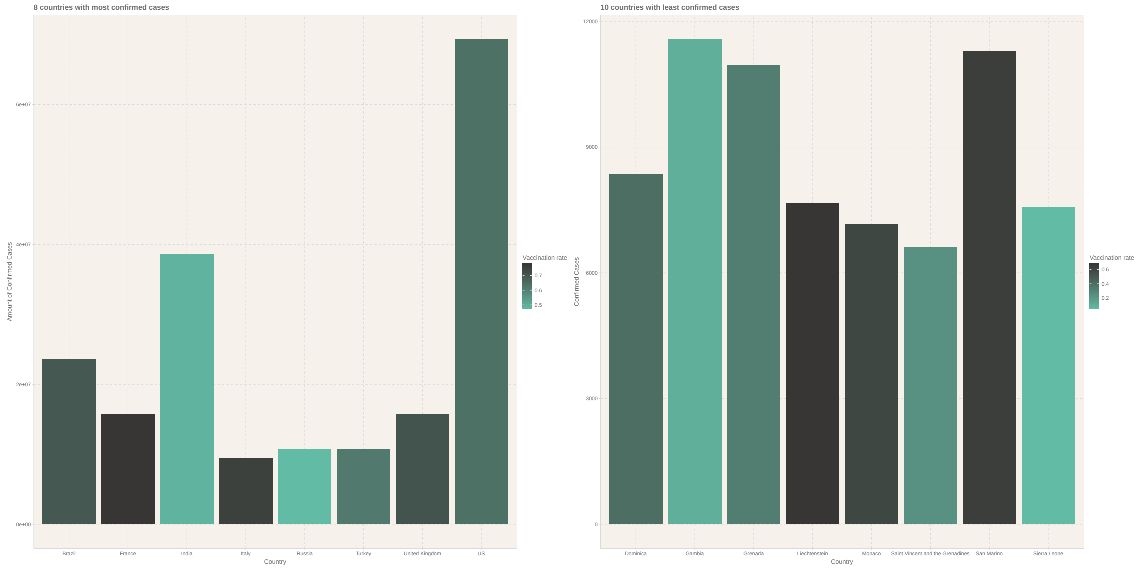


Figure 4 Countries with the most/least confirmed cases and their vaccination progress

graph is generated analogously, only the palette has to be changed if it supports less than 9 colours. The extended visualisation can be seen in figure 5. The exact division of the categories is of course adaptable and should be adjusted in case of doubt depending on the domain knowledge. In the chart, the vaccination rate cut-offs 10%,30%,50%,60%,70%,80% and 90% were used.

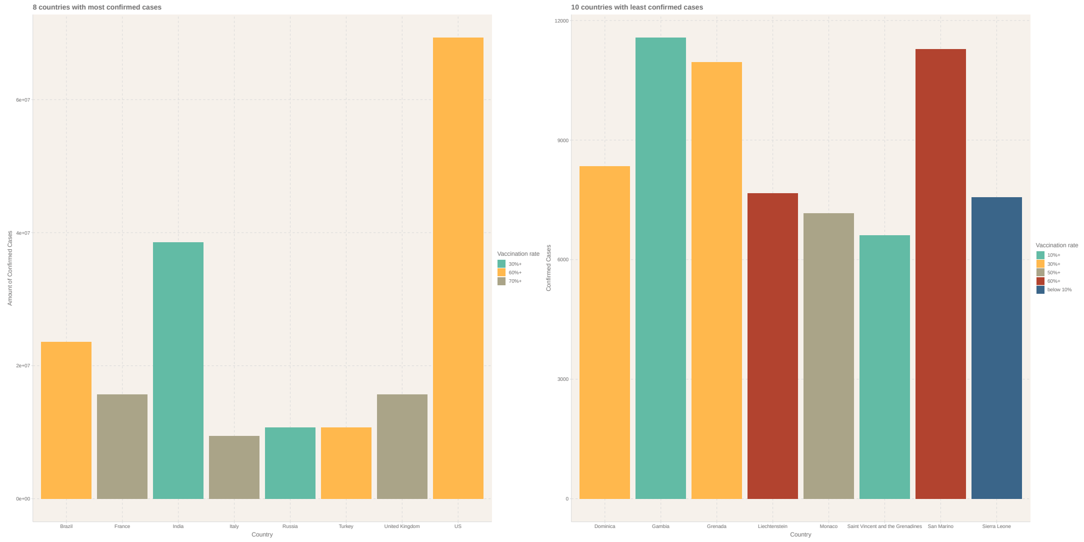
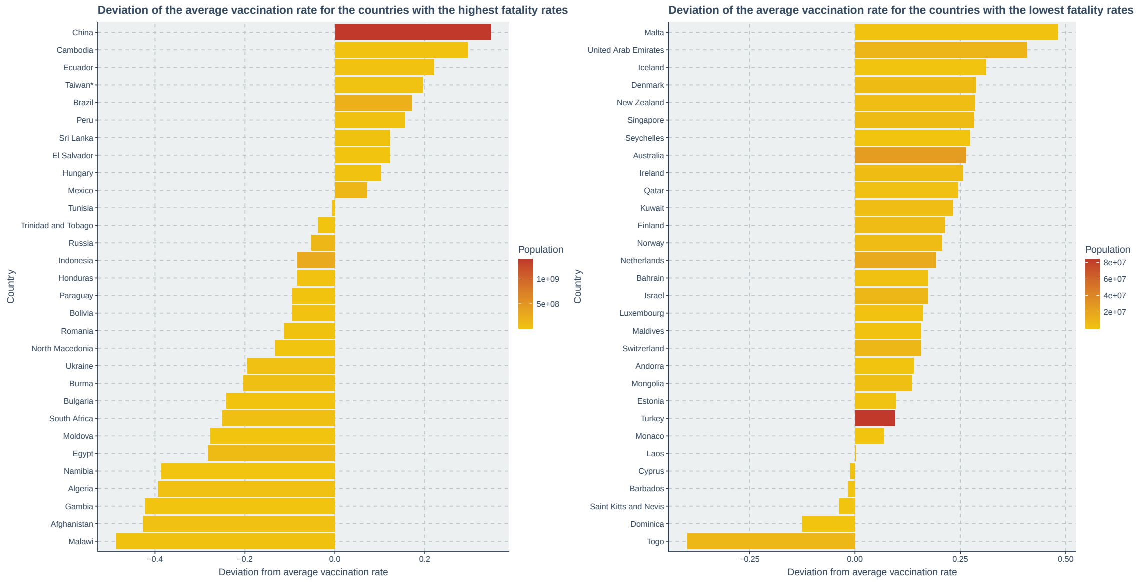


Figure 5 Countries with the most/least confirmed cases and their vaccination progress (categorical)

4.3.2 Deviation of the mean vaccination rate depending on fatality rate

In the second chart (cf.. figure 6), I want to contrast a country's Corona fatality rate with its vaccination progress. To do this, we use a bar chart centred on the mean value of the global vaccination rate. To see demographic differences directly in the context of the population, the filling of the bars is mapped with the population data. In each case, the countries with the highest/lowest lethality rate were selected. The countries are sorted by this number in descending/ascending order.

To visualise this, the data set is first prepared appropriately. First, we filter out the data sets without fatality rate and vaccination rate. The order of the countries is determined by the fatality rate. To calculate the deviation, we first calculate the mean value of the vaccination rate and then subtract it from each vaccination rate. Thus, the data are now centred on this point and can be plotted in a bar chart. The population can be recognised by the filling. In order to recognise finer differences, the theme is varied this time. It is noticeable that a large part of the countries with high fatality rates have a below-average vaccination rate. The opposite is true for countries with low fatality rates.



Deviation of the average vaccination rate for the countries with the highest fatality rates

Deviation of the average vaccination rate for the countries with the lowest fatality rates

Figure 6 Countries with the most/least confirmed cases and their vaccination progress

4.3.3 Worldwide vaccination progress

In the previous visualisations, mainly cases at the edge of the definition range were analysed. One could almost speak of extreme cases. In view of the pandemic, however, it is also interesting to know how the world as a whole is performing. Therefore, the following chart (cf. figure 7) analyses the vaccination status of the world, detached from specific countries. The vaccination progress is again recorded as a category and now shown in a treemap, so that one can directly see what proportion of the world's countries are vaccinated and to what extent. Such a chart could, for example, play a strategic role for measures. The exact basic quantities can be adjusted again. This could also be created for the federal states of Germany. Other variables could also be displayed as sub-hierarchies.

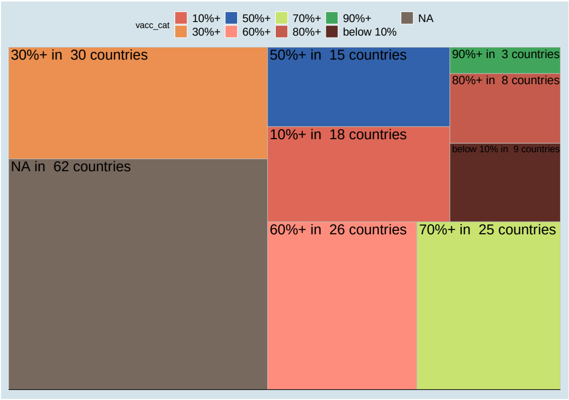


Figure 7 Aggregated Vaccination progress of the world

4.3.4 A comparison of case/death figures and the vaccination rate

With the help of a bubble chart, the possible (positive) effects of vaccination are now presented. For this purpose, we compare the vaccination rate with the confirmed case numbers or the death figures. This forms a scatterplot. In addition, however, we map the diameter of the points with the population of the respective country and thus finally form a bubble chart. You can see in the chart that the smaller bubbles are deeper and there were fewer corona cases there. This can be reasonably explained by the different population figures, as these are absolute numbers. In addition, one can see that from a vaccination rate of about 75% onwards, there are clearly more data points in the lower range, i.e. lower case and death figures.

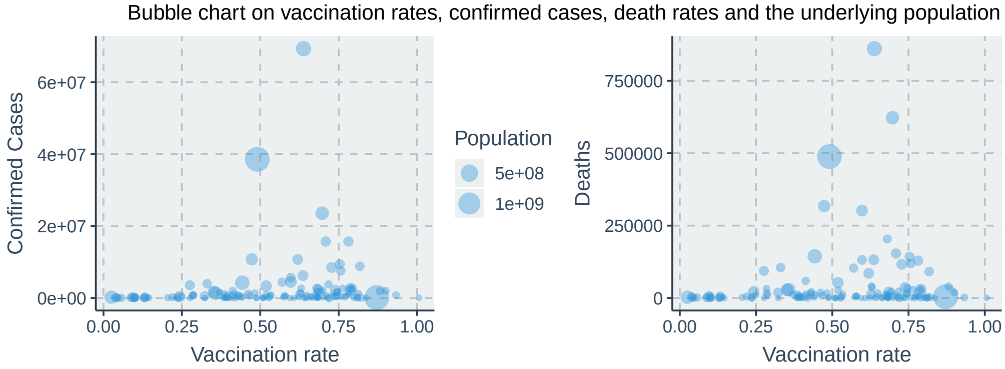


Figure 8 Comparison of case/death figures and the vaccination rate by using a bubble chart

4.3.5 Time series visualisation of confirmed cases and vaccination numbers

Finally, we look at the visualisation of time series. To keep the chart (cf. figure 9) clear, only the DACH region is considered. In addition, a trend line can also be displayed. You can see here that Germany has the most cases and vaccinations in the DACH region. The values for Austria and Switzerland are similar. After a longer period of time, Switzerland now has significantly more confirmed corona cases. This increase occurred recently and is therefore hardly visible in the trend. Austria has been vaccinating more doses than Switzerland for a long time. Liechtenstein's figures are hardly discernible due to the low numbers. This can be explained by the significantly smaller population.

In order to generate the chart, data preparation had to be carried out again. The data set was first restructured (melt-function) so that each data point contains the date. Previously, each date was a single variable and the data set was therefore very broad. After this step, the variable had to be parsed from the date so that R could recognise the date in order to display it. Similarly, the vaccination data had to be processed. Both charts were plotted under each other.

Since the data preparation was carried out universally, one could now visualise the course for any country as desired. If you want to make forecasts, you can add relevant key figures to the chart, such as a moving average forecast. If you have seasonal data available, or suspect seasonality, you can also include this. For example, you can remove the trend from the chart. If you want to take a closer look at the trend, you can create a time-series-decomposition and display the long-term trend, the seasonal trend and the noise separately.

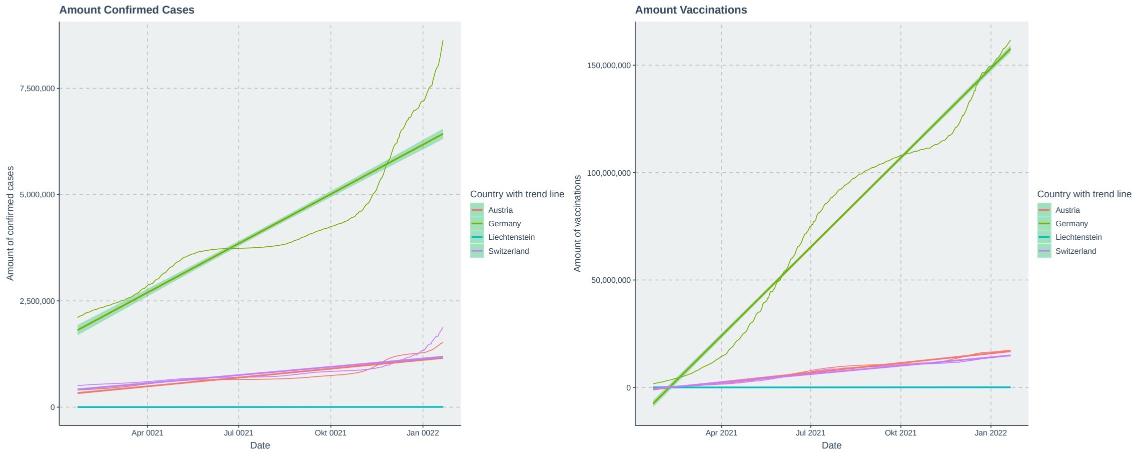


Figure 9 Corona time series analysis of the DACH rergion

4.4 Evaluation of the visualisations

The uncertainty of the Corona visualisations has already been addressed. Before the results are analysed, short goals and requirements are defined. In doing so, we orientate ourselves on positions from the literature.

Visualisations serve to help with analysis, understanding and communication. This can refer to models, concepts and data (cf. Schumann and Müller 2000, p. 14).

In our context, the communication of data is addressed, as this is usually done with Corona visualisations in everyday life. So it is about communicating the statistics in graphics. There are three levels that a good visualisation can achieve. We speak of the first level when the basic information is presented in the graphic without further restrictions. The second level is reached when the graphic supports the reader and clearly presents the result of the investigation. The last stage is reached when the entire data set is shown with all hidden information. Such a visualisation could be used without hesitation for important decisions, such as the definition of corona measures by the state (according to Schumann and Müller 2000, pp. 15-17).

In general, it should be noted that it is hardly possible to create visualisations of the third stage within the framework of this work. The data sources are not absolutely correct and there is great uncertainty in the pandemic.

The first graph was primarily concerned with case numbers and used immunisation statistics as supporting information to help understand and situate. This indicator was used particularly heavily at the beginning of the pandemic, although it shows clear weaknesses. The absolute number of cases is not relevant and it is therefore normal that countries with a higher population score "worst" in this statistic. It is also not surprising that small countries such as Liechtenstein appear in the list. At this point, data literacy would be required from the reader. In addition, the author may have deliberately set the minimum threshold at one level in order to influence the selection of countries. Without background, one could conclude that the countries on the right-hand part are particularly well-suited to the pandemic.

A continuous indication of the inoculation value may be difficult to read, so a discrete indication with categories works well. The graph helps the reader to get an overview of the situation. The differences in size are very easy to see and one quickly understands what this visualisation is about. Thus, the first two steps have been reached. The third level is not reached, because the population is not reached. In addition, a sample was taken by the author, so it should not be used as a decision-making aid.

The second graph also serves its purpose. It tells the reader specifically what the position of the countries with the highest/lowest fatality rate is in terms of vaccination coverage. The centring helps the reader to quickly get to grips with the graphic and the subject matter. In combination with the colouring, the essentials quickly become clear. The second stage is thus fulfilled.

Nevertheless, the graph must be interpreted with caution. The situations in the countries are different. There are demographic, economic and social differences. The population alone is not sufficient to explain the result. Other relevant information such as the local infrastructure or medical situation is not known. However, the graph is well suited to give a small overview of the statistics. Despite the overview, this does not necessarily mean that one may conclude causality.

The third graph shows the global vaccination campaigns of the different countries and has divided them into different categories. The colours and area sizes quickly show the difference in size. The uncertainty also becomes clear, as many countries do not have correct data. Through the aggregation, the goal is almost completely fulfilled and everything is displayed. The first two stages are fulfilled. However, as long as countries have not provided any data, the population is not covered. Stage three is therefore not achieved. Nevertheless, the graph serves the purpose exceptionally well and answers the question about global immunisation success very well. The segmentation of the categorisation could be enlarged in order to increase the amount of information. At this point, a trade-off has to be made. The visualization is closer to the third level than its predecessors because it uses the information from the data source.

The fourth graph compared vaccination rates with case rates and death rates. Primarily, the communication of the data is important here, since the quality of the data source is not high enough to draw firm conclusions. The information is easy to understand. However, data literacy is required because the reader must be aware that many more factors play a role in this question, population alone is not sufficient as a segmentation criterion. Without data literacy interpreting might be problematic. Analogous to the previous visualizations, level 3 cannot be reached either.

In the last visualization, the time series is analysed of the number of cases and vaccinations. Since in this chart we were only interested in the DACH countries, the data from the dataset is fully taken into account and no information is lost, as in the first two charts. The question about the progression can be answered quite clearly and precisely. For stage three, one would have to ensure correct calculation and processing of the data. So, as a consequence, it comes very close to the definition of the third stage.

5 Conclusion

In summary, visualizing data can be a very challenging activity, not only during a pandemic

When visualizing, there is a lot of theoretical background that goes into the work. Many of these already happen subconsciously, as the author recognizes when a visualisation is not optimal. Nevertheless, these theoretical backgrounds must be taken into account. As learned in the context of perception, small details, such as a type of form, can result in large changes in perception. It is imperative to pay attention to this. The colour choice is a good example here, since this often becomes problematic as soon as many variables are in play. In the course of this work, the colour palette was therefore changed at necessary points or presented in a supplementary variant. The choice of a good colour is also responsible for the first correct impression. For the latter, the aspect of preattentive search is relevant. If this concept is ignored, the reader may not be able to recognize the essential information or may have to search for a long time. This would contradict the second step of Schumann and Müller (2000).

In the second chapter the special relevance of the visualisation of the corona statistics and the accompanying problems were discussed. The pandemic determines the everyday life of all of us day by day. At the moment, even after two years, it is impossible to imagine life without it. In times of corona measures such as lockdowns, the results of statistics have a direct impact on the way one can lead one's life. This is a distinguishing feature of the statistics of the Corona Pandemic. This is not true for every statistic people encounter in everyday life.

Also important is the serious aspect of uncertainty in the statistics, since different institutions are involved and the state of knowledge is not final.

In the fourth chapter, visualisation, it becomes clear that data visualisation is not just about creating graphics. The main work lies in the conception and data preparation. This often involves working with different data sources. These do not necessarily have to be coherent with each other. Although the data of this project are provided by the same institution, they were not coordinated with each other. Therefore, it is again emphasized that the conception is of extreme importance. Before visualisation, the data set was first analysed manually in detail until it could be understood.

During the visualisation process, it was realized that different graphics could be used to help describe a situation. Five extensive visualisations were created with the R package ggplot2, which required individual preparation and cleaning.

To conclude the work, the visualisation were analysed for their purpose using a concept from science. With the result that the visualisation often cover the questions very well. Nevertheless, it happens with the first two visualisation, for example, that the population is not considered. In the background one must think back here to the aspect of the uncertainty.

References

Frank, M., Walker, J., Attard, J., and Tygel, A. 2016. “Data Literacy - What Is It and How Can We Make It Happen?” *The Journal of Community Informatics* (12:3).

Graham, C. C. 1869. “What Is Perception?” in *The True Philosophy of Mind*, Unknown Publisher, pp. 131–134.

Healy, K. 2018. *Data Visualization: A Practical Introduction*, Princeton University Press: Princeton, United States of America.

Loss, J., Boklage, E., Jordan, S., Jenny, M. A., Weishaar, H., and El Bcheraoui, C. 2021. “Risikokommunikation bei der Eindämmung der COVID-19-Pandemie: Herausforderungen und Erfolg versprechende Ansätze,” *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* (64:3), pp. 294–303.

Schliffka, C. 2021. “Verlässliche Daten auch in Krisenzeiten – die deutsche EU-Ratspräsidentschaft im Bereich Statistik während der Corona-Pandemie,” *WISTA – Wirtschaft und Statistik* (73:3), pp. 28–35.

Schumann, H., and Müller, W. 2000. *Visualisierung*, Springer Berlin Heidelberg: Berlin, Heidelberg.

Timcke, M.-L., and Schneider, B. 2021. “Welt aus Daten Datenjournalismus während der Corona-Pandemie,” *Zeitschrift für Medienwissenschaft* (13:25-2), pp. 102–114.