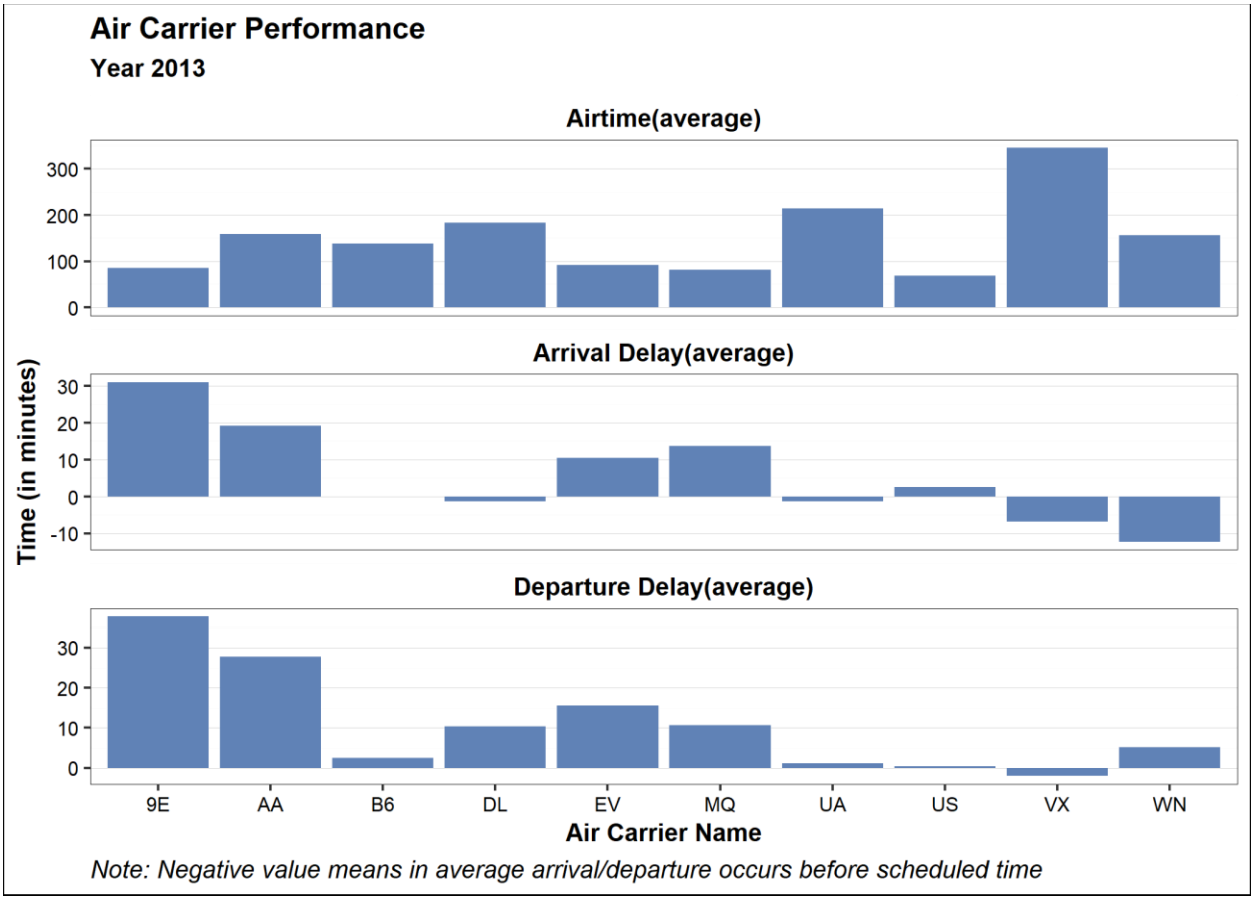
Part 1. Graphic inquisition

Source of the figure: https://www.statisticshowto.com/wp-content/uploads/2014/01/Fox_welfare-jobs-ff-1.jpg

My findings from above figure of Fox News are given below -

1. This graph does not comply with the Gestalt principles regarding simplicity and figure-ground. Presence of wheel shaped objects in the ground takes more attention than the figure.
2. There is scope to work with the operations of this figure. Title for X and Y axis, grid line and scale on Y axis are missing. Color used in bars and the numbers seem like to be different layers and makes the figure somewhat haphazard that is difficult to decode.
3. This figure can be a perfect example of why maximizing the data-ink and erasing non data ink is important. The wheel shaped objects that have been unnecessarily used here might distract or impede the viewer's interpretation of the data. Also, the strong bar colors, color of background and number box colors have made it look like someone has put different layers of things together to make this figure. Title size is also unnecessarily bigger in ratio to figure.
4. This figure is hugely criticized for its graphical data integrity and lie factor. According to source, Y axis includes every individual residing in a household in which one or more people received benefits in the fourth quarter of 2011. On the other hand, X axis includes only individuals who worked, not individuals residing in a household where at least one person works. Thus, they have shown a larger number of people who were on welfare than the number of people who had jobs. Beside this, I can add, they have shown bigger effect in figure than the data. The actual effect is somewhat 6% in data, but in figure the bar is almost 4 times bigger. Seems like they have started their scale in Y axis from 100.
5. This figure could be easier to read alone with title of X axis and Y axis, proper grid line and proper annotation.

Part 2. Graphic Design



In this figure I have shown average time of different parameters of Air Carriers that can provide a quick view of the performances of those carriers. This graph can provide a quick view of the carrier performance to the passengers so that they can plan their flights accordingly. It can also be used by the governing bodies to improve the performance or utilize the carriers which are less used or reduce the airtime of the carriers that are overly used.

Average airtime: It provides the average time the carriers spent in travels. Governing bodies can use this to use these carriers in an efficient way.

Average arrival delay: It shows the average arrival delay in minutes. Passengers can get help from this and plan their journey accordingly. Managing team can also use this to improve the timing. Negative value indicates in average arrival occurs before scheduled arrival time.

Average departure delay: It shows the average departure delay in minutes. Passengers can get help from this and plan their journey accordingly. Managing team can also use this to improve the timing. Negative value indicates in average departure occurs before scheduled departure time.

As I have criticized a bar plot designed by Fox News, I have tried to incorporate all the 5 criteria of good graphs in my designed bar plot.

Part 1. Graphic inquisition

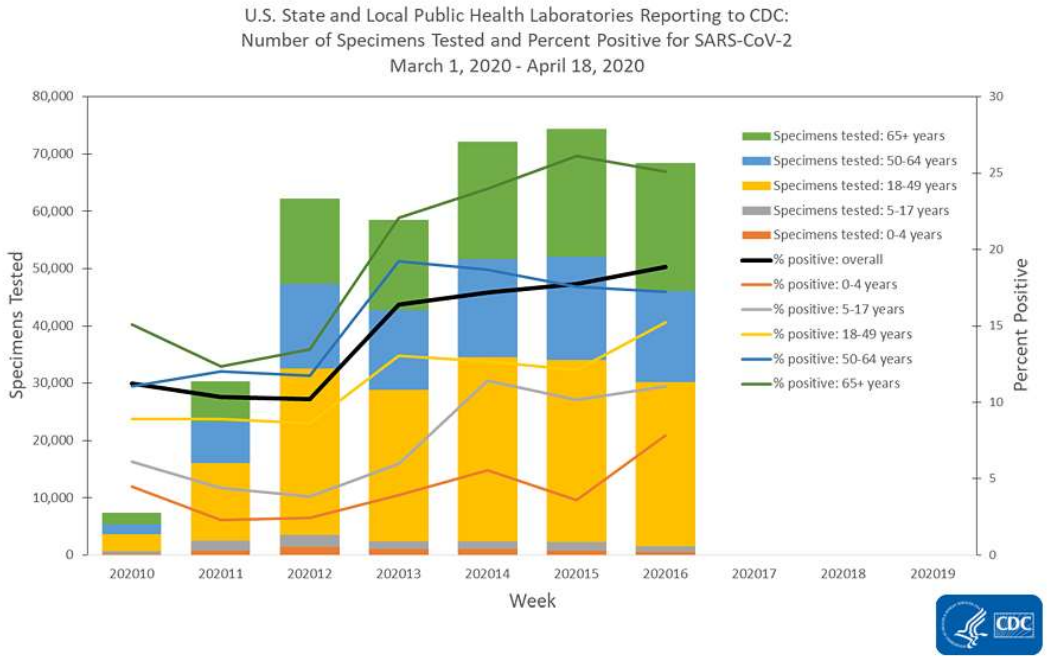


Figure 1: The graph by the Centers for Disease Control and Prevention (CDC) website created based on the weekly data provided by U.S. State and Public Health Laboratories during the surge of Covid 19 in March/April 2020.

Source: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/04242020/public-health-lab.html

Last checked online- 18/10/2022 at 3:27.

The above CDC graphic was intended to show the number of specimens tested and percent of positive cases for SARS-CoV-2 in one single graphic by plotting both the variables in double Y- axis where X-axis depicted the age groups.

The graphic looks quite decent. Using stacked bars and simple lines makes it visually structured. Color is contextual and meaningful -five contrast colors represented 5 different age groups. Black line is explicit for readers to understand overall positive cases

The graphic, at the first sight looks a bit complex, but it is not super-hard to decode the necessary information. However, the uneven (jiggling) base line along X-axis makes the bar diagram heavy to understand the trends for the 65+ (green), 50-64 (blue), and 18-49 (yellow) age groups. Grid lines could be used to better anchor the reference points for effective scanning of bar compositions. Superimposition of bar and line graphs made it reader unfriendly.

Indeed, no chartjunk, distracting or redundant element was used, and data-ink ratio is fair enough to display the data on top of everything.

The graphic is trustworthy since the data is reliable and accessible on CDC website. However, on X-axis different age groups contain variable age differences, such as 4 (0-4), 12 (5-17), 31 (18-49) and so on, which might distort the size effect and confuse the readers.

Unclear and ambiguous labelling of year and week on X-axis makes the readers puzzled for a while. Even the last three points remained incomplete with no data shown but the legends lied upon. Double Y-axis were used to indicate the bar and line charts making it hard for viewers to immediately identify the values of individual data points. The coordination between the line chart and corresponding Y-axis (right) is lost in the space occupied by the legends.

Cramming together two quantitative variables into a single, overly-complex graph made it difficult for readability on its own with minimal effort. Rather, small multiples visualization concept introduced by Edward Tufte could be a good choice to make at least two vertically or horizontally faceted charts for better stand-alone readability.

Part 2. Graphic design

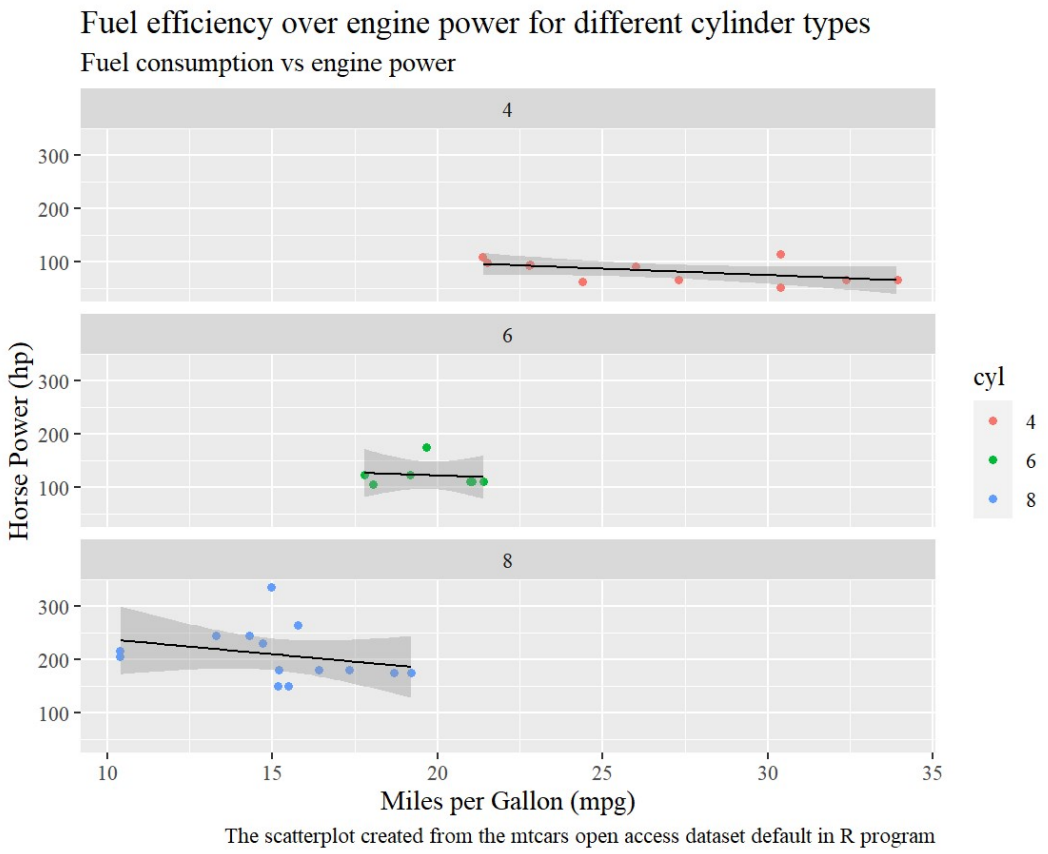
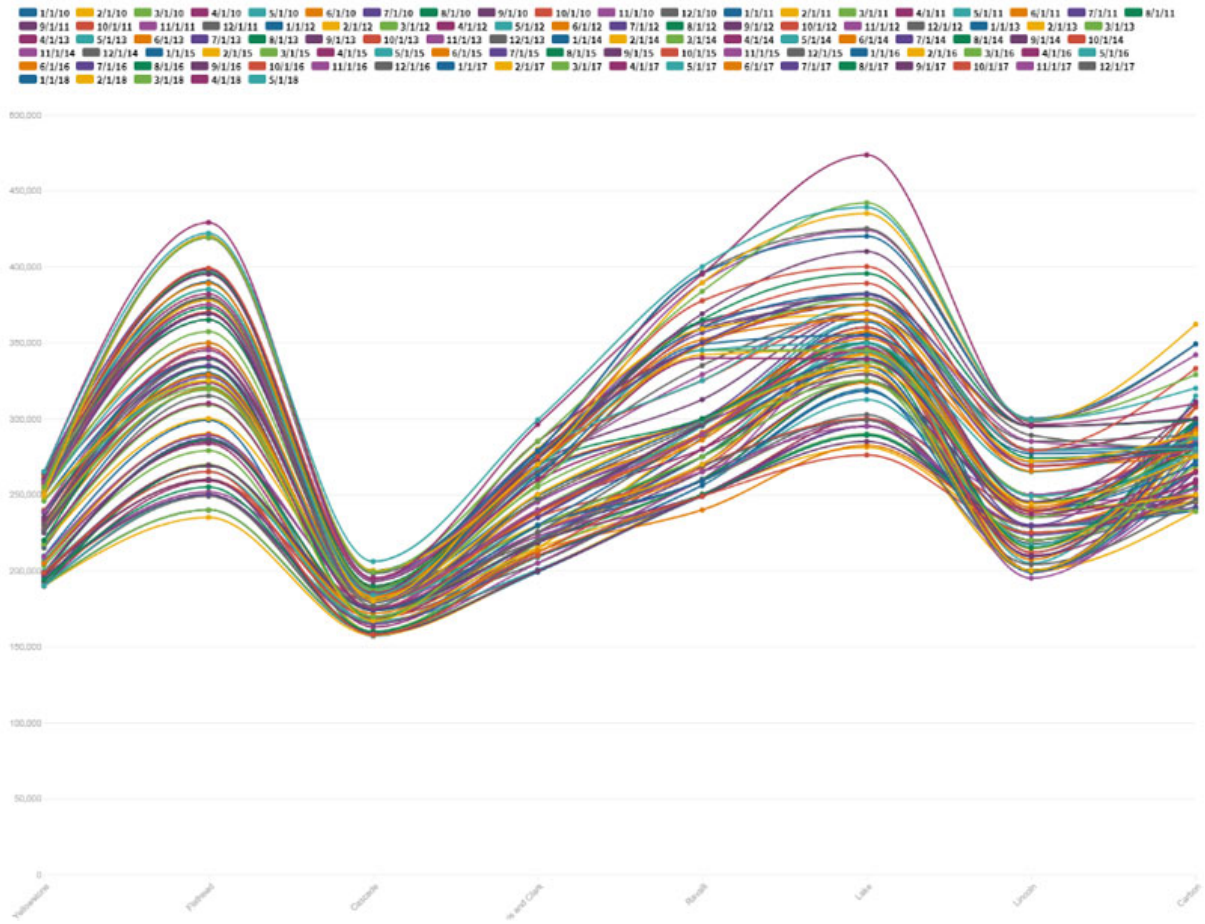


Figure 2: How fuel efficiency varies with the engine output for mtcars with three different types of cylinders.

The mtcars dataset contains measurements on 11 different attributes for 32 different cars. A car engine with higher cylinder numbers would be more powerful and ideally would consume more fuel. Hence, the plot was intended to see how fuel efficiency (mpg) varies with the engines' output (hp) in case of different cylinder types (cyl).

For the graphic, ggplot2 was chosen because of its advantages options over the basic graphics, such as adding multiple layers, themes, facets etc. making the visualization structural, effusive and more engaging. Two discrete (hp & cyl) and 1 continuous (mpg) variables can better be represented by scatterplot where "cyl" was converted to categorical for differentiating facets. The "geom_ point()" could sufficiently show the relationship between two aesthetically mapped variables ("mpg", & "hp"), however. "geom_jitter()" is better to show even small amount of random variation to the location of each point, and to avoid overplotting caused by discreteness in this smaller datasets. The layer "geom_smooth()" was used to aid seeing patterns in the presence of potential overplotting and method "lm" was used to fit linear models. Labels was helpful to make the plot interactive. The panel facet_wrap() was well-enough to form the matrix by single variable "cyl".

Three facets show the relationship between fuel consumption and engine power for 3 types of cylinder cars. In case of 4-cylinder, fuel efficiency reduces from 34 to 22 mpg where engine power increases from 52 to 120 hp indicating a negative linear relationship. Negative relation though weaker is observed between mpg and hp for 8-cylinder cars. In contrast, the 6-cylinder cars are prone to have less impact on fuel efficiency (fluctuating around 20 mpg) while engine power increases from 105 to 175 hp. Overall cylinder number has inverse proportionality with the fuel efficiency meaning high cylinder number car has low fuel efficiency and vice-versa. The graphic showing some data points remain outside the geometric smooth linear model indicating discreteness – that might be due to impact of other constant and latent variables (such as disp, wt, gear.).



Link: https://wpdatatables.com/misleading-data-visualization-examples/

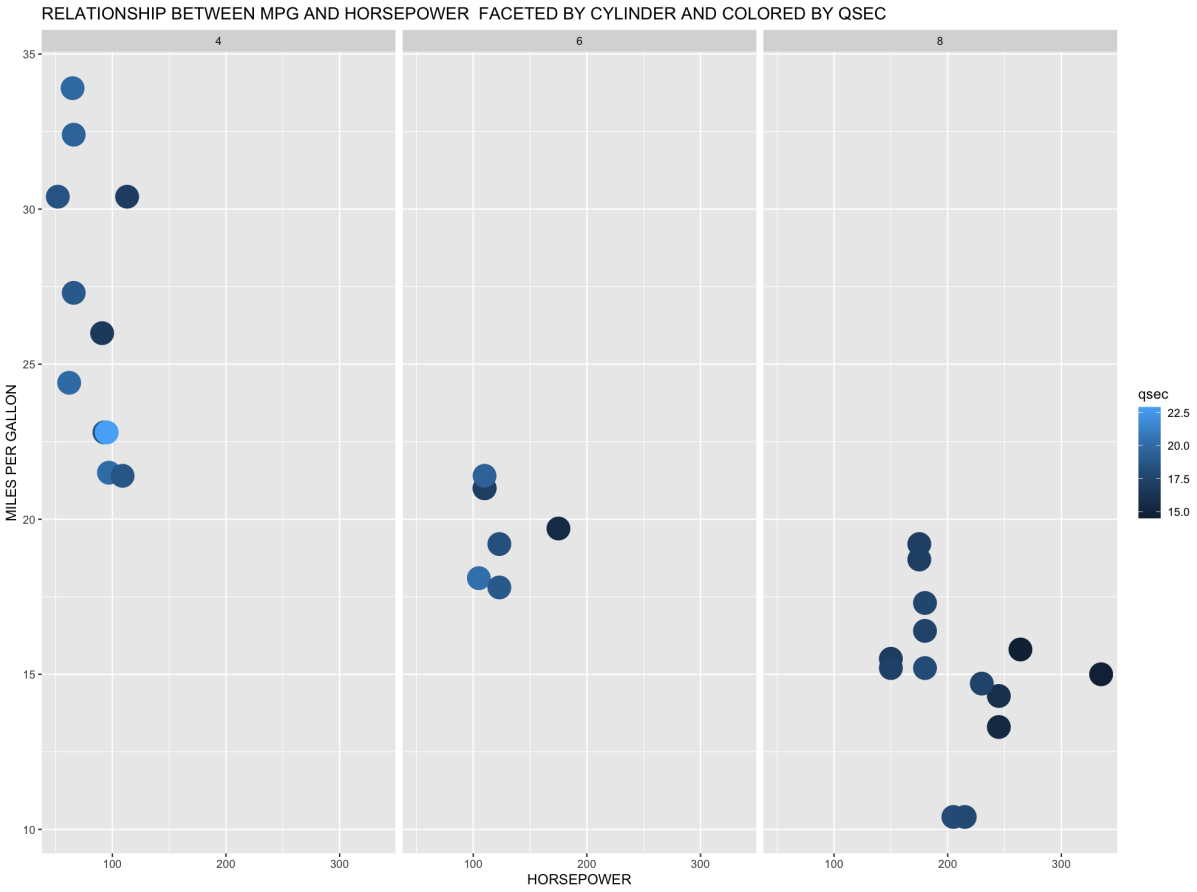
It is usually frustrating when one had a lot to communicate to a non-technical audience within a limited time period. Though graphs serve an important purpose, sometimes, people make misleading statements using graphs. The attached graph represents a bad representation of graph. It shows the number and range of exam students over a long period. It showed that universities admitted students from minority groups and lower income families.

To start with, this graph failed the gestalt principle and visual structure. Some colors are similar. To be honest, one must also show clarity with graphs. Otherwise, it will be a cos 90 job since the message might not be understood. With regards to the Keep it simple criteria, this graph once again clearly fails it. There are too many variables, and this makes it difficult to differentiate between data points. When there is a lot of data, there is also a lot of interesting details, and this makes decision making harder. The best way is to know what data format is the most effective way of communicating a point clearly.

Another principle which has been failed is the Less is more principle. This is like the keep it simple criteria. This graph looks like a chart junk and it's very difficult to differentiate between data points. Also, a person with color issues might find it difficult to differentiate between them as there are so many confusing colors here. One must strain the eyes in order to see the labeling on the X and Y axis. This makes it hard to read.

Also, the graphical integrity and lie factor has been failed. One only wonders if too many data points and variables were used in order to obscure certain data points. Lastly, the annotation of the graph is difficult to read. The X and Y axis are so faint that one must strain the eyes just to see what is written there. Also, the annotation on top of the graph which shows which color represents which date has so many similar colors which makes it hard to distinguish. Stand-alone readability is impossible in this case as it so difficult to distinguish between data points. There are so many data points that makes it impossible to differentiate between them.

To conclude, this graph is a true definition of a bad graph as it has failed all the five criteria.


RELATIONSHIP BETWEEN MPG AND HORSEPOWER FACETED BY CYLINDER AND COLORED BY QSEC
4
6
8
MILES PER GALLON
35
30
25
20
15
10
100
200
300
HORSEPOWER
qsec
22.5
20.0
17.5
15.0

RELATIONSHIP BETWEEN MILES PER GALLON (MPG) AND HORSEPOWER.

I wanted to show the relationship between and the horsepower of a car. To do this, I facetted the cars by using the number of Cylinders. This resulted in 3 Facets (4,6 and 8 cylinders). Also, I also colored each data point using the qsec ¼ mile time.

Horsepower refers to the power an engine produces. There is a negative relationship between horsepower and miles per gallon. Cars with low miles per gallon have higher mpg has is shown in the graph. The data has been split by the number of cylinders. Cars with 6 cylinders have a higher horsepower than cars with 4 cylinders and cars with 8 cylinders also have a higher horsepower than cars with 6 cylinders. The higher the number of cylinders, the higher the horsepower. This is clearly depicted by the graph as each facet represents the number of cylinders.

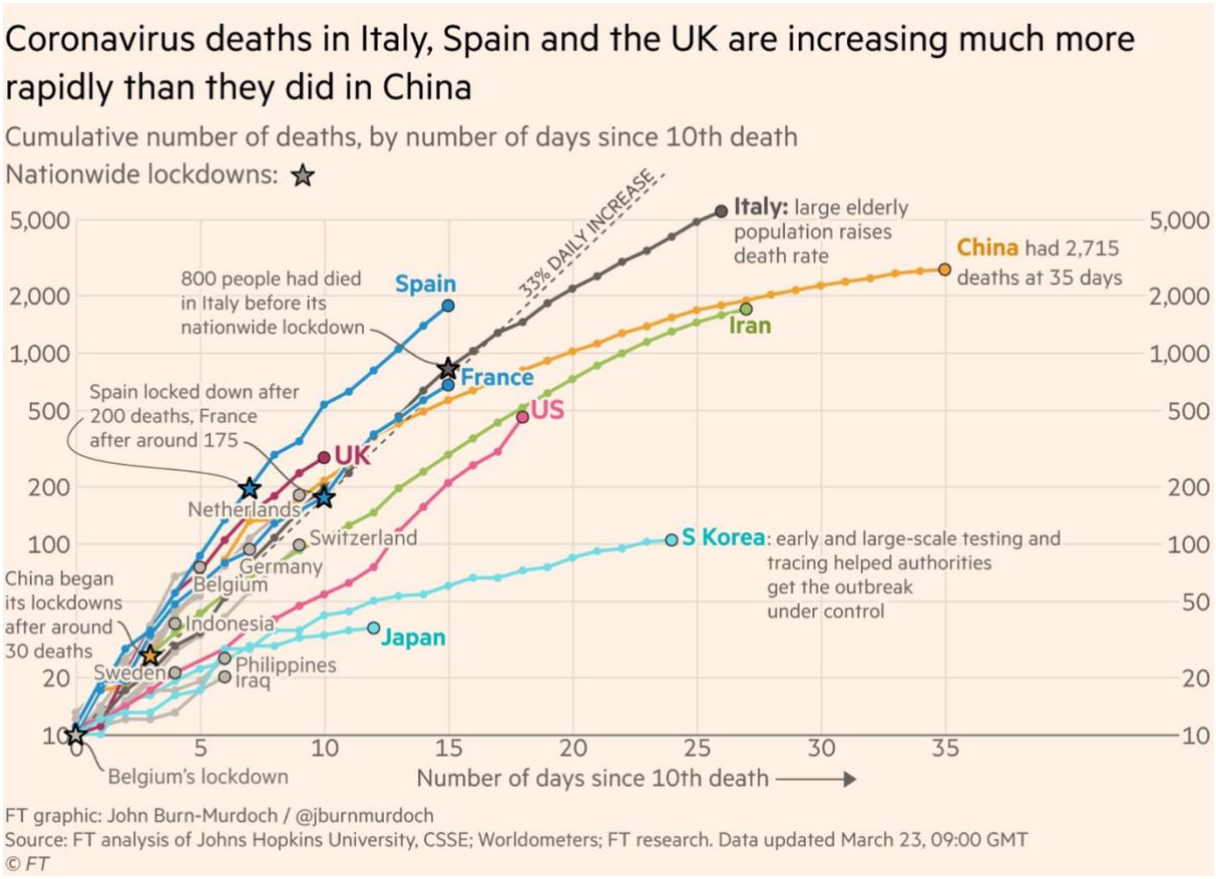
Also, the data points are colored by the acceleration (qsec) which is ¼ mile time. Thus, the acceleration it takes to cover ¼ mile. As can be seen, cars with a higher horsepower usually accelerate faster and covers the distance easily. The darker the color, the faster the car. Cars with 8 cylinders have a darker color the cars with 6 and 4 cylinders. And cars with 6 cylinders also have darker colors than those with 4 cylinders. Thus, darker colors use an acceleration of 15.0 to cover a quarter mile. The lighter colors use an acceleration of 22.5 to cover a quarter mile.

The colors in my graph are bit similar and it might be difficult for someone to differentiate between them. Adding different colors might make it easier for people to understand. Also, there are too many variables, and this can make it difficult to interpret.

Part 1. Graphic inquisition

Figure 1

Coronavirus deaths in Italy, Spain and the UK are increasing much more rapidly than they did in China



Note. John Burn-Murdoch [@jburnmurdoch]. (2020, March 23). You asked, we answered: The @FinancialTimes coronavirus death & case trajectory trackers are now 🔥 FREE TO READ 🔥 outside the paywall: Http://ft.com/coronavirus-latest In this morning's update, the US has gone above 470 deaths, bringing it just behind where Iran was at the same stage Https://t.co/NBA7FMYlmC [Tweet]. Twitter. https://twitter.com/jburnmurdoch/status/1242048518908502016

A relatively simple graph gets messy with all the information about the other countries that are not in focus. It is hard to see which parts of the graph correlates to the stated objective of the graphic. Proximity of the lines, possibly due to the interwall choice makes it hard to read. Similar colors e.g., Spain & France makes it difficult to track each nations specific progress.

The position of specific points on the lines is hard to see, especially between x=5 and y=50. Bad color saturation e.g., Japan/S Korea & Spain/France, but it is good between the four *main* observations (UK, US, Spain, Italy). Decent use of grid lines. The scale is consistent, and the dual y-axis is helpful. Reader is not informed of the log-scale. Decent projection, some superimposing is needed in the bottom-left of the graph.

Decent data-ink-ratio. Some unnecessary comments on nations such as S Korea, France, and Belgium. Notes on when lockdowns were enforced are not necessary due to the star annotation that is stated under the title. It could be said that since the only four countries that we want to observe are UK, US, Spain, and Italy, all other should be removed. Good use of grid lines, easy to differentiate between grid lines and countries in grey.

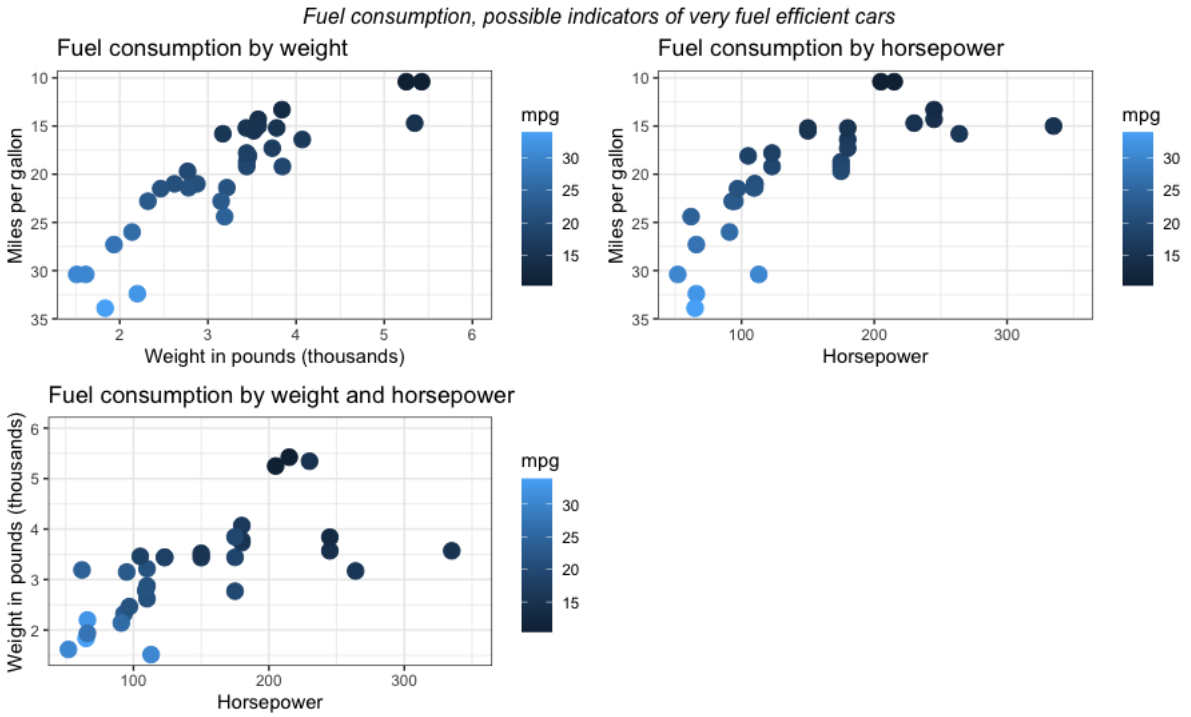
Good graphical integrity, but the y-axis is a log-scale, and this means that the scenarios that look similar, like China and S Kora are far from equally comparable, even though they look so similar. Due to the non-specified log-scale, the graph is not in comply with this factor.

Unnecessary comments decrease readability e.g., comments on lockdowns and comments on countries other than those in the title. No indication of log-scale, can´t see “50” on the left y-axis. Annotations on the graphic are somewhat informative. It is difficult to immediately find specific datapoints, but with some time you can interpret the graphic by itself. Good stand-alone readability. The claim that covid is more rapidly increasing in Italy and Spain, compared to China is clear, but we cannot say the same for the UK.

Part 2. Graphic design

Figure 2

Fuel consumption, possible indicators of very fuel efficient cars



The graph intends both to educate and explore which common measures of a car has the most impact on the car being very fuel efficient. When looking for a car, two easily available statistics are the weight of the car and the amount of horsepower it produces.

In panel 1 we can see that there is a clear correlation between the weight of a car and its fuel consumption. The higher the weight, the less distance it can travel per gallon of fuel it consumes. The panel shows us a general trend, but there are around the 2000-lbs-mark several cars with quite different mpg. On the other side of the spectrum, the three cars above 5000 lbs have a low mpg, but the variance between them are still noteworthily high. Based upon this, we can conclude that while weight can give us a general indication about cars fuel consumption, there seems to be other differences that also play a large part in the calculation.

Looking at panel 2 we can see that a high amount of horsepower correlates to having low fuel consumption. This gives us a general understanding of horsepower being a factor in the cars fuel efficiency. The variance in mpg compared to horsepower is very large and there are cars with a low amount of horse power that have both a very low, and quite high fuel consumption. As with the conclusion in panel 1, this shows us that horsepower alone cannot determine to a significant level whether or not a car is fuel efficient or not. But contradictory to panel 1, horsepower seems to be a worse indicator than weight.

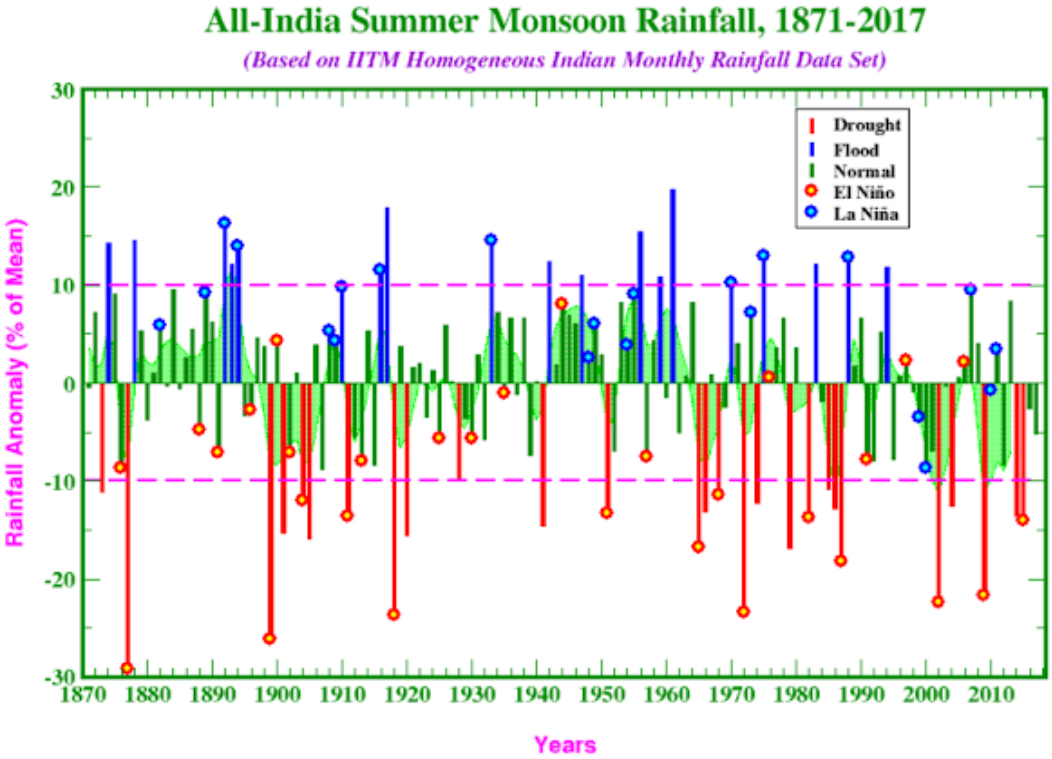
Panel 3 explores the fuel efficiency of cars by looking at both the cars horsepower and its weight. This would give us an insight into weather a combination of both variables could give us a good indication on whether or not a car is very fuel efficient. While this produces a somewhat generalizable trend, there are both cars with low weight and high horsepower, and high weight and low horsepower that have similar mpg. A combination of both variables does not seem to give us any new insights other than both weight and horsepower having an impact on the cars fuel efficiency.

Component II: Data Visualization

Part 1: Graphic Inquisition

Figure 1.

Graphic for Inquisition



Note. Graphic obtained from Macromyths blog authored by Srinivas Thiruvadanthai (2018). It is unclear whether Thiruvadanthai is the author of the graph or if he is only referencing it.

Discussion

Figure 1 was obtained from a blog by Thiruvadanthai (2018), and it depicts the rainfall in India from 1871–2017. It also seems to aim for the inclusion of other relevant weather phenomena (El Niño/La Niña, drought, flood, and average rainfall). The figure is difficult to understand and read, and including so many variables does not serve its objective.

Gestalt Principles and Visual Structure. The figure is not characterized by simplicity since multiple elements like lines, area, data points, and colors coexist. Other Gestalt Principles are acceptable such as figure-ground differentiation and symmetry through a balance of elements.

Decoding and Operations. The figure uses the most effective elements for decoding, such as position in a scale and length, but then makes heavy use of the least effective: color—a lot of color. Anchoring through a grid is omitted, which helps keep some simplicity to it, but apart from that, the figure has too many superimposed elements that make its processing difficult in the end. Superimposing was a poor choice for this data.

Chartjunk and Data-Ink Ratio. There is some chartjunk and excess ink for the data presented in this figure: the drought/flood lines could be unnecessary and seem invasive. Color inclusion is indiscriminate and plain chartjunk. Axes are duplicated in the opposed borers of the plot, and minor breaks are unnecessary. There is a green area that is unclear which variable it represents.

Graphical Data Integrity and Lie Factor. The graphical integrity is not too bad, as there is no evidence of severe disproportionality to deceive the reader.

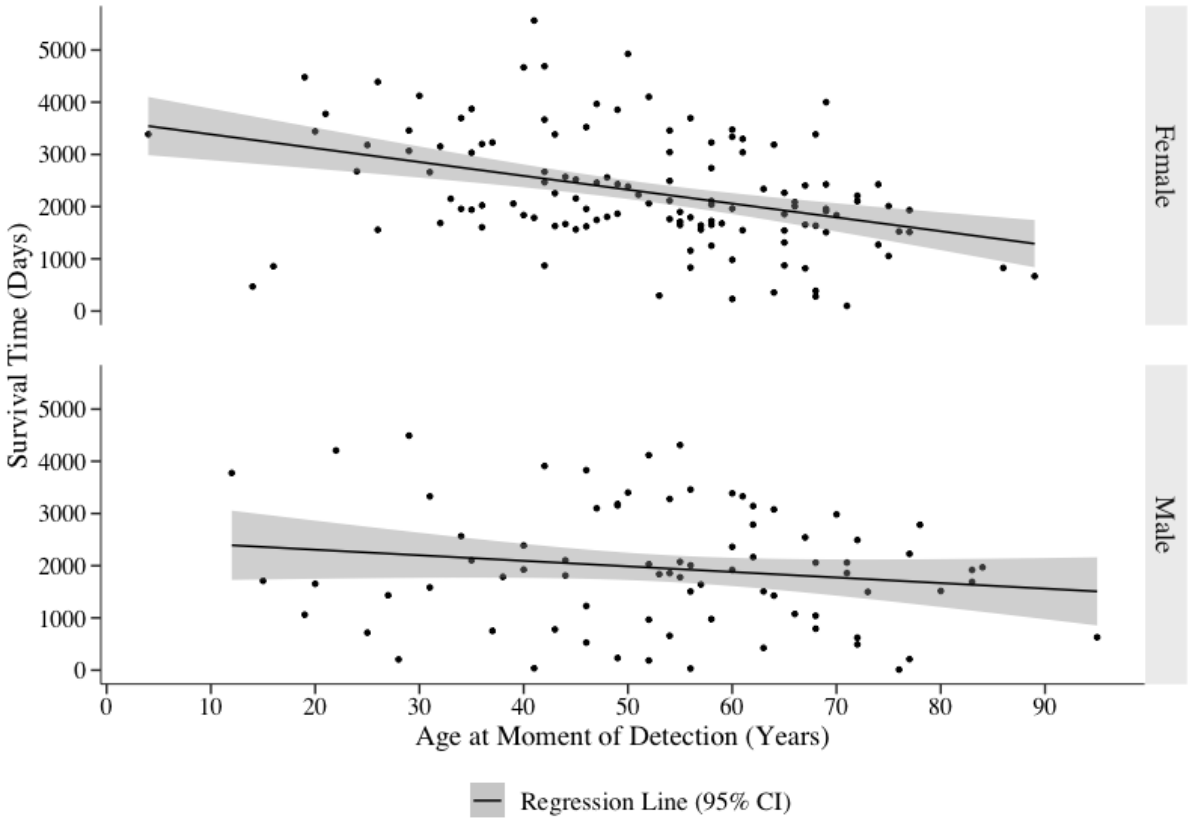
Annotation and Stand-Alone Readability. The figure can be a stand-alone read, but this is at the expense of a lot of text and information in the plot itself, so this item could be considered as failed. The annotation does not follow the APA style complicating its readability.

Part 2: Graphic Design

Designed Figure

Figure 2.

Scatter Plot Representing the Relationship Between the Age at Melanoma Detection and Survival Time



Note. Data obtained from the "Melanoma" dataset which is included in MASS R package (Venables & Ripley, 2022).

Discussion

Figure 2 is a scatter plot based on the "Melanoma" dataset from the MASS package, containing data on 205 patients with malignant melanoma in Denmark. The chosen variables were patient age in years (continuous), days of survival (continuous), and gender (binary). With the figure, I wanted to account for the potential relationship between age and survival time and compare this among gender. I opted for a scatter plot since both variables were continuous, and the data set had a relatively significant number of data points.

Gestalt Principles and Visual Structure. The designed figure complies with the Gestalt Principles. The plot is simple and clear. Both facets are symmetrical and can be read with continuity. Data points and lines are simple and similar across the figure. Proximity and similarity between the panels help to understand that they come from the same sample.

Decoding and Operations. To ease the decoding of the figure, both panels use common scales and share one axis. Color was omitted, and saturation was used only when strictly necessary. Regarding operations, grid lines were omitted as the objective was to present a general idea of the relationship rather than locate every data point. Scanning is eased by presenting consistent breaks in both survival time scales. To compare by gender, projection was chosen instead of superimposing to ease the graphic's processing.

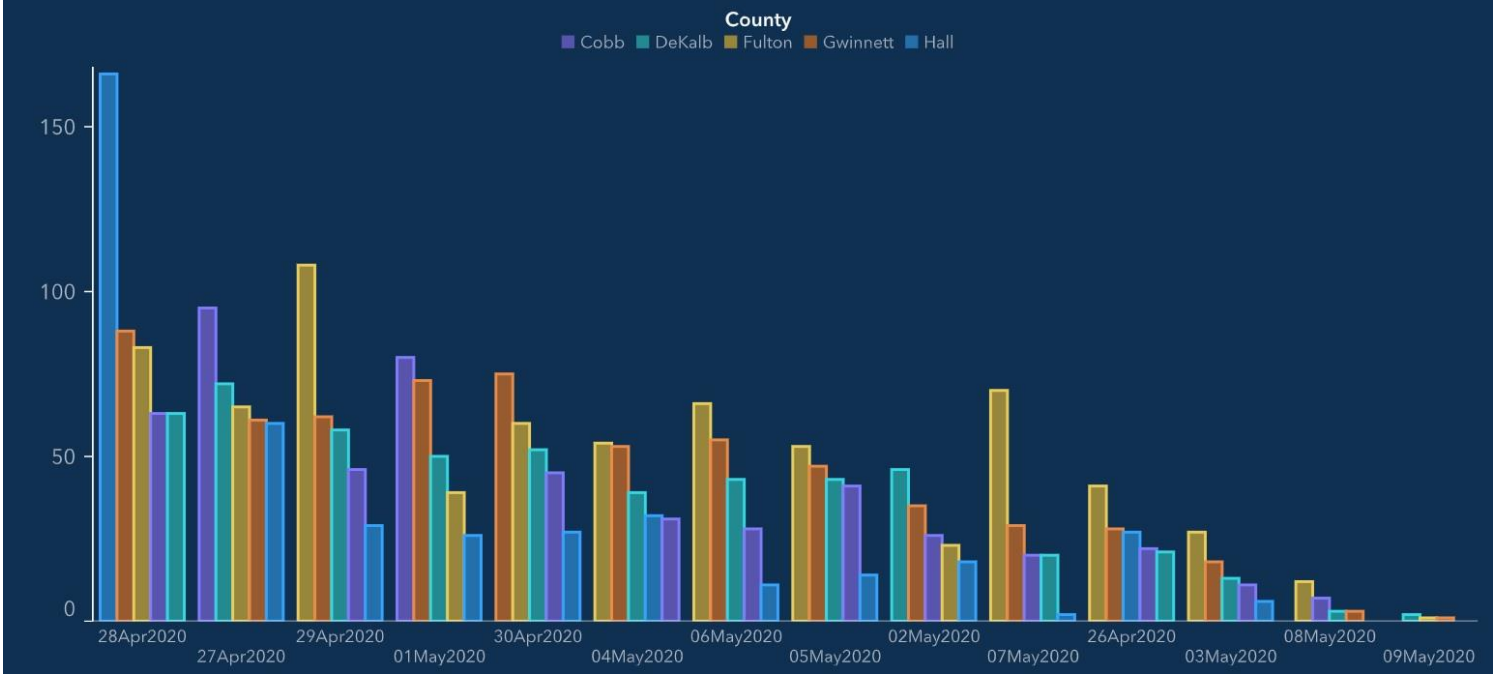
Chartjunk and Data-Ink Ratio. The graph contains only strictly necessary information. Grid was omitted as it would not add information. No chartjunk elements are present.

Graphical Data Integrity and Lie Factor. Labels are consistent and meaningful, and the effect size of the data and graph are consistent and not exaggerated in any direction. The aspect ratio of the axes helps preserve the data's integrity.

Annotation and Stand-Alone Readability. The figure is readable by itself and is self-explanatory as it contains all the information for its understanding. It is annotated and formatted according to APA reference style.

Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



The graph I have chosen is collected from vox.com and is supposed to represent the top five counties with the greatest number of confirmed COVID-19 Cases. There are multiple reasons why this is not a good graphic representation. The graph itself is simple enough at first glance, without too many unnecessary design choices included. It's a simple bar graph with values on the y-axis and dates on the x-axis with bars representing each of the counties. Even though it's easy enough to figure out in which country these counties are, it should also be included in the graph to make it completely clear what its giving information on. The graph shows a decrease in cases of deaths and hospitalizations by regions.

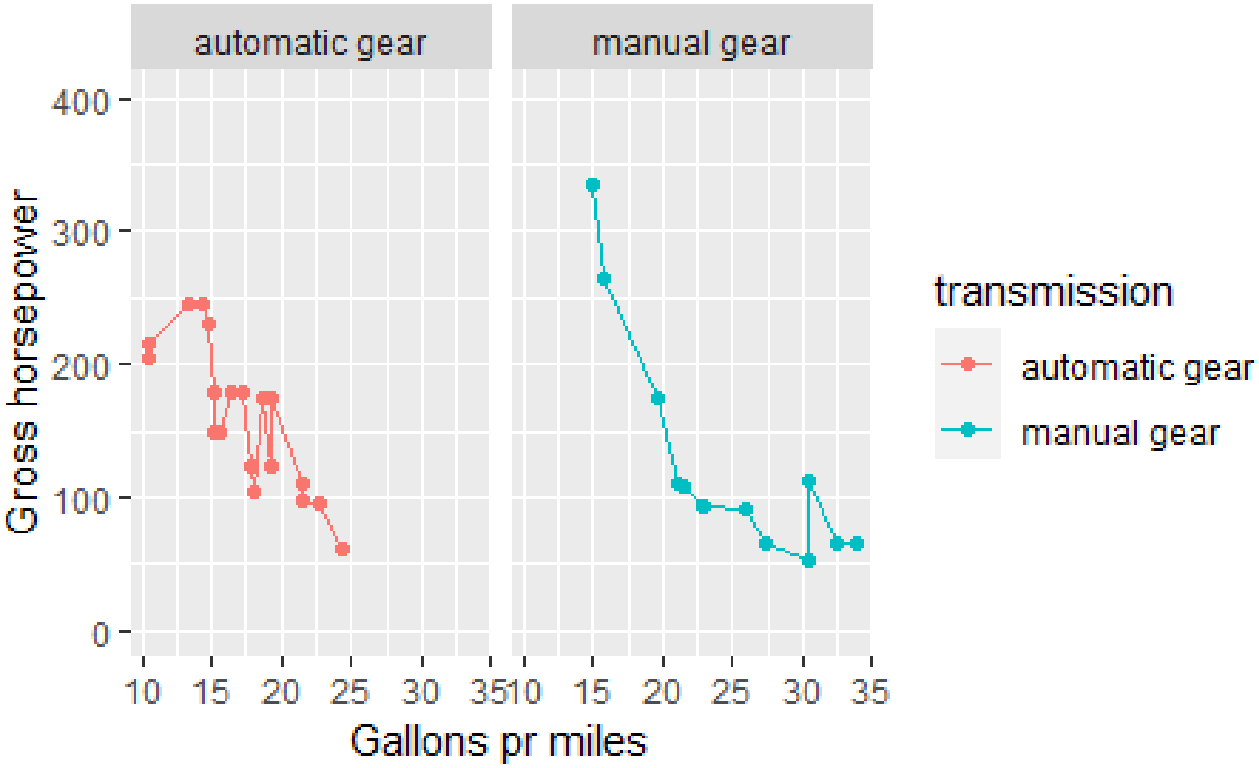
Neither of the axis's are named though. For the x-axis it isn't that big of an issue since it's clear that its dates. On the y-axis on the other hand, it's difficult to understand what the scale is. In the info text below the headline, they claim to show both the number of deaths and the number of hospitalizations. We can therefore assume that's what the numbers are supposed to represent, even though it's possible to understand, it should be made clear in the graph to avoid misinterpretation.

Another big issue with this graph is the order of the dates on the x-axis not being chronological. You will have to read the graph very thoroughly to be able to understand what the order of the dates are, and then being able to visualize how the graph is supposed to look based on that.

The last issue the graph is facing is the randomly changing order of the counties within each date. The creators have made the bars decreasing in size within each of the dates, which gives a wrongful presentation of the evolution within each county.

Bibliography

Collins, S.(2020,may 18th). *Georgia's Covid-19 cases aren't declining as quickly as initial data suggested they were".* VOX. *https://www.vox.com/covid-19-coronavirus-us-response-trump/2020/5/18/21262265/georgia-covid-19-cases-declining-reopening?fbclid=IwAR3DB_sTbO0Vcf9c8MBbBhO43bsHj1bEAgt7ov2q1-3WAS7QxLJpNd1KPPE*


automatic gear
manual gear
Gross horsepower
400
300
200
100
0
10
15
20
25
30
35
Gallons pr miles
transmission
automatic gear
manual gear

The graph shows how a car's horsepower affects its usage of fuel and is divided by transmission type. The horizontal axis, or the x-axis, gives us a scale of the usage of fuel from 10 to 35. The vertical axis, or the y-axis, gives us a scale of a car's horsepower from 0 to 400. This is a choice made because to clearly show that there are a hundred horsepower between each bar. The limit is given because the maximum horsepower a car is documented in the data to have, is a little under 350, and to keep the consistency of a hundred between each number, I set the limit to 400. This is also to avoid any lie-factor and have the size of effect shown in the graphic true to the size of effect in the original data.

The first facet, showing only cars with automatic gear, shows us that there is a positive correlation between a car's horsepower and the usage of fuel. The second facet, showing only cars with manual gear, is giving the same results where the scores start higher than with automatic geared cars. This can tell us that cars with manual gear do in general have a higher level of horsepower than the ones with automatic gear.

I have chosen to keep the graph simple when it comes to layout. I have tried to stick to the gestalt principles when it comes to simplicity to avoid any confusion regarding what the graph intends to show. The colors are only applied to the plots, in order of easing the readability of the graph. I changed the name of the facet variable from am to transmission to avoid any confusion. All in all, I focused on the “less is more” principle and kept the graph as simple as possible, while still including everything necessary.

[Portfolio] Component II: Data Visualization

Figure 1

Graphic inquisition

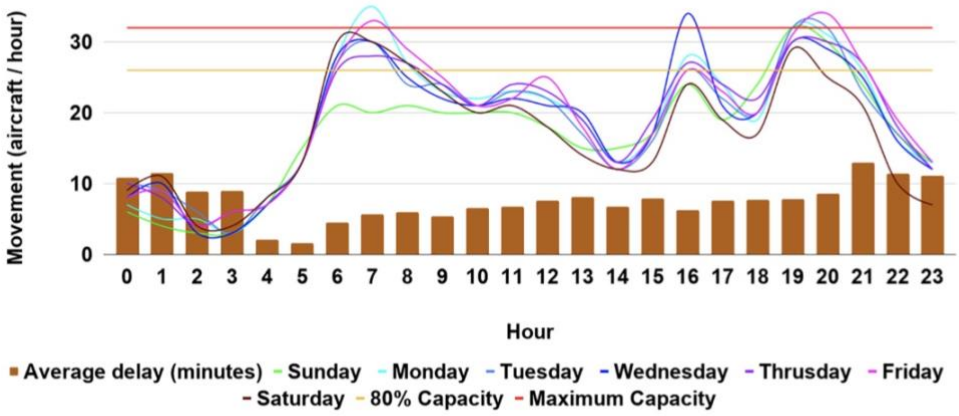


Fig. 9. Maximum actual demand, airport capacity, and average delay at SBGR in 2017.

Note: This image was retrieved from Pamplona and Alves (2020).

Pamplona, D. A., & Alves, C. J. P. (2020). An overview of air delay: A case study of the Brazilian scenario. *Transportation Research Interdisciplinary Perspectives*, *7*, 100189. https://doi.org/10.1016/j.trip.2020.100189

Graphic inquisition

Gestalt Principles: In the figure (Figure 1) the authors make use of the Gestalt principle, *common fate*. The demand variable for different days is moving in the same direction, so we perceive these as part of the same group. Same principle for the bars measuring delays in minutes.

Decoding and operations: There are grid lines in place to help the viewer decode the values to do further operations. The element lines, used for plotting symbols, are however on top of each making it hard to tell them apart. The color scale also makes decoding hard. The y-axis isn't scaled properly ending at 30 when the value of *Maximum capacity* (red line), and most of the demand variables, indicate higher values than that for some hours of the day. On the y-axis, they're using the same values for two different units. This gets confusing when decoding the values from the bars in the plot because you have one element that's easier to decode (lines) along with an element that's harder to decode (bars) on a common scale measuring in different units.

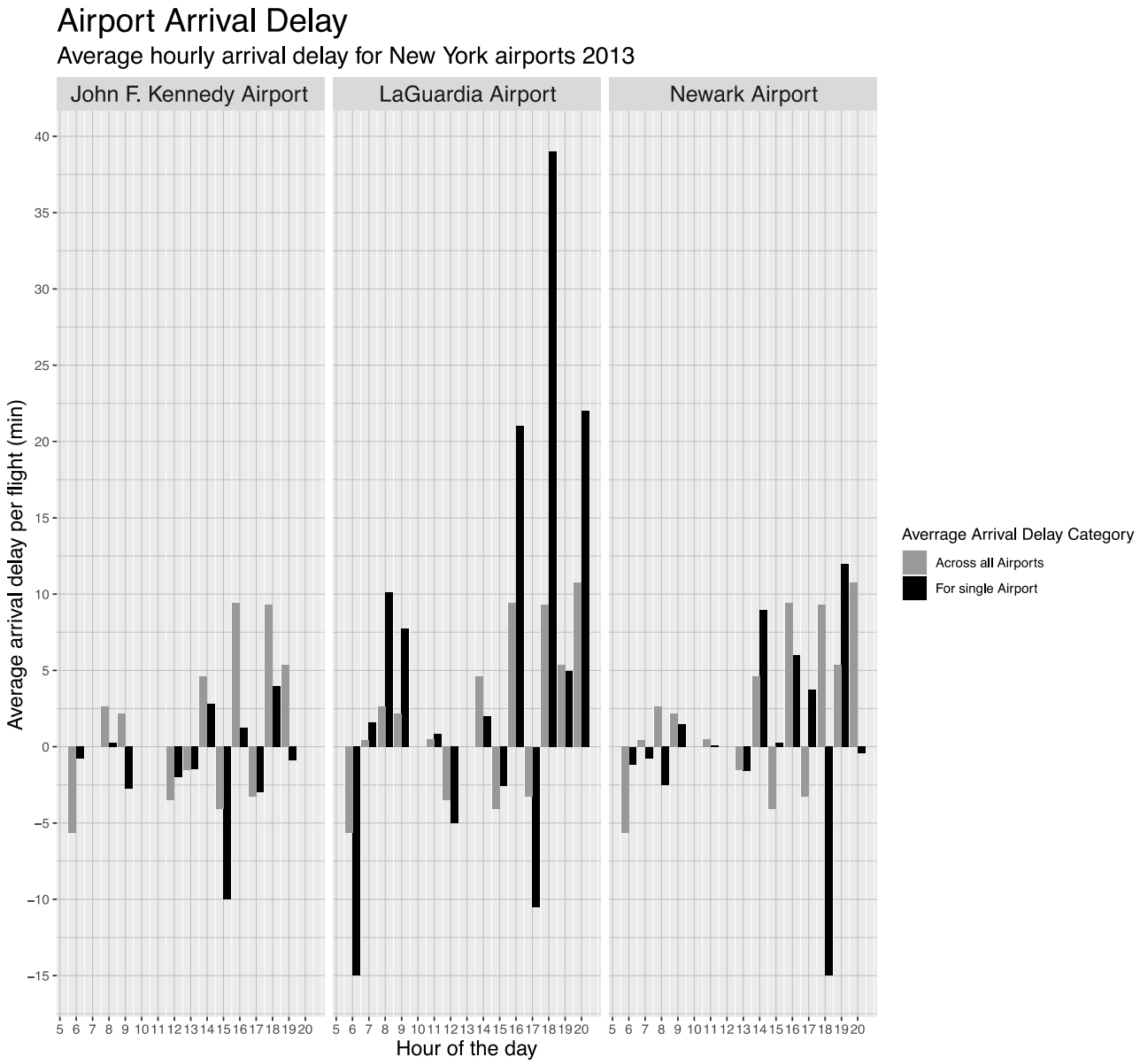
Annotation and stand-alone readability: Most of the variables are marked. The figure title (Figure 1) also provides enough information to be interpreted outside its context (the article). Preferably, the *Maximum Capacity variable* could've been named the same thing in the figure title as it is in the notes.

Graphical data integrity and lie factor: It doesn't seem as if the author has misrepresented their data. The x-axis contains all hours of the day, not creating large increases/decreases vertically. It seems as if they've adhered to the 2H:3L aspect ratio.

Chart junk: The data-to-ink ratio is quite low. Because they include many variables in one panel, they use a lot of colors (=ink). Also, they don't clarify the use of the 80% capacity variable. They might mention a reason in the article, but as a viewer, it's not clear. Also, if they'd scaled the y-axis properly, I would've been able to read the value of the maximum capacity, making computation of the 80 % of the maximum capacity easy.

Figure 2

Graphic design



Note: This graph is generated using a dataset with 150 entries for the year 2013. Outliers in the data have also been removed. An outlier is defined as being 1,5 * Inner quartile away from the 2nd and 1st quantile. After doing that, the remaining number of entries were Newark (58) John F. Kennedy (47), and LaGuardia (34). The average arrival delay per flight is computed over the number of flights remaining after outliers were removed. No data was available for the 10th hour, and for each airport where no flights for giving hour were given, there's also no Average Across all Airports displayed. Data named "flights.txt" was downloaded from the UiO canvas folder.

Graphic design

I chose to display the average arrival delay per flight for each origin airport. I also chose to display the average arrival delay across all airports for comparison.

Decoding and operations: I chose to include gridlines to easily derive values from the figure for further computation. And I put the bars I wish to be compared next to each other so that the viewer doesn't have to move objects mentally, referencing the term, *Superimposition*. I chose a color palette that would be accessible even to those with color blindness.

Annotation and stand-alone readability: I labeled each panel, gave the figure a title, and labeled the legend so that the average category was clear. I noted in the figure text how the data was processed before being visualized.

Graphical data integrity and lie factor: I chose to create a variable mean arrival delay per flight to be able to compare the three airports. One could argue that it's not fair to LaGuardia having the least departing flights and so one single delay, excluding the outliers had a massive effect. But then again, this is not a representation of reality anyway. I note that the number of entries from which the mean was calculated for each airport is in the figure notes (Figure 2).

Gestalt principles: I chose to include the mean for all airports in each of the panels. I think this helps the viewer put the values for each panel into perspective, and I also use the same plotting symbols for all averages conveying that they're symbolizing the same unit (arrival delay per flight (min). Adhering to the gestalt principle, *similarity.*

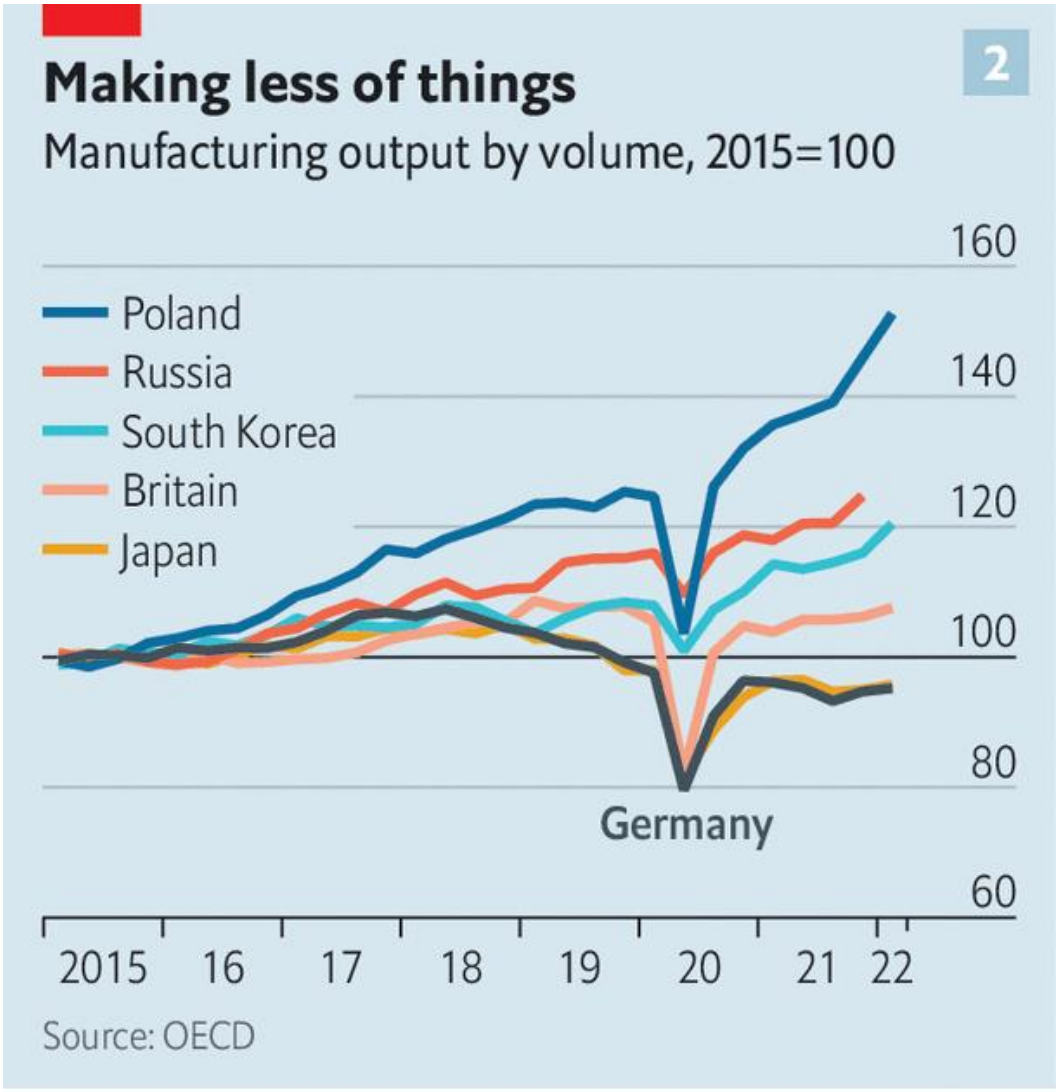
Chart junk: I debated including the average across all airports because the viewer can calculate this using the graph. But I kept it, thinking it visualizes a comparison directly to the viewer.

Visualization component

Part 1. Graphic Inquisition

Figure 1

Making less of things: Manufacturing output by volume, 2015=100



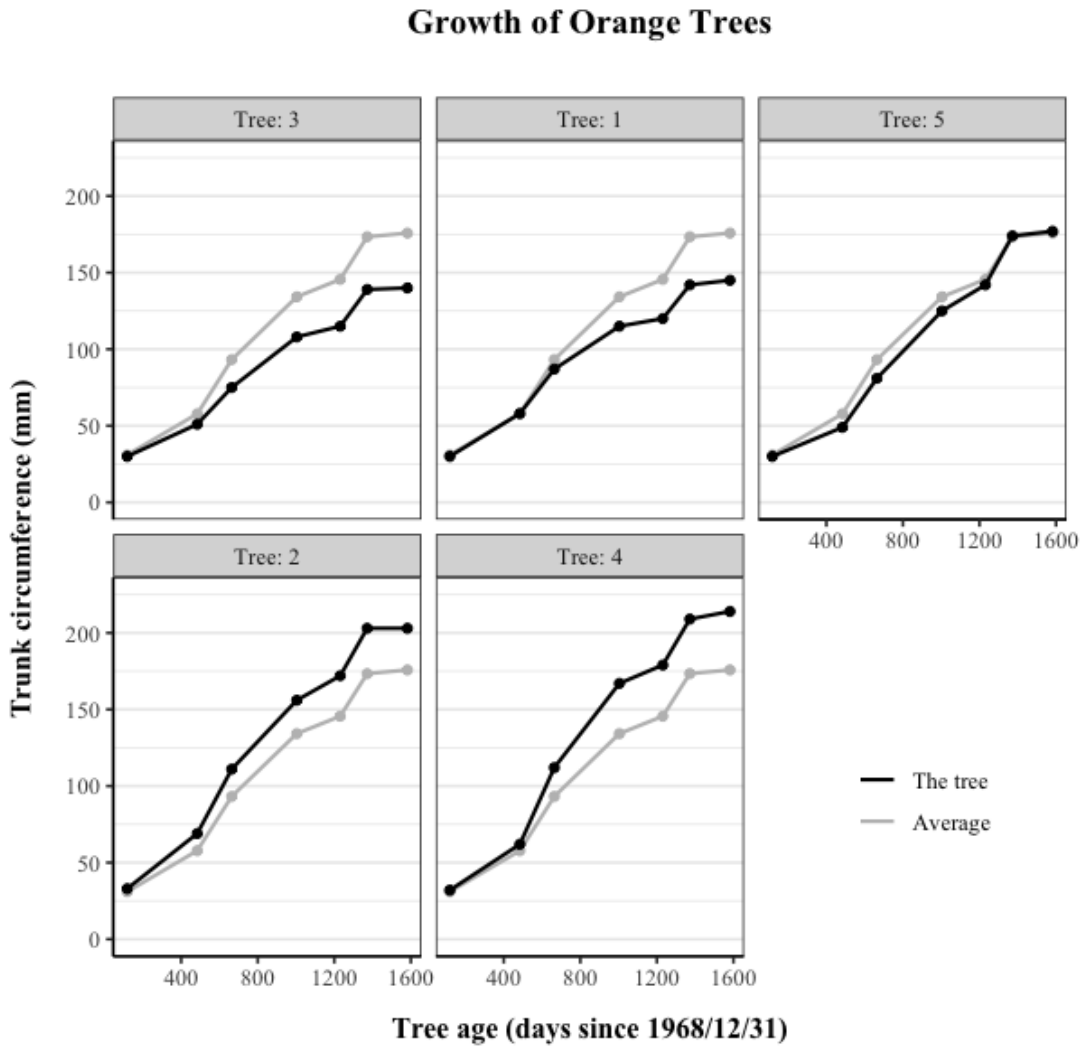
Note. The Economist. (2022). Schafft Deutschland das? Retrieved October 06, 2022, from https://web.archive.org/web/20220811214418/https://www.economist.com/briefing/2022/08/11/germany-is-facing-dramatic-change-in-many-dimensions-all-at-once

*1. Keep it simple:*The figure intended to show that "German manufacturing is no longer growing in absolute terms", according to the source article and the figure title. However, this intention is not that easy to be perceived by the readers. Since there are too many colors in one plot, and the colors may not easily be distinguished by everyone. We may take more effort into distinguishing and comparing different color lines. *2. Gestalt principles & visual structure.* The colorful lines overlap a lot at the left end of the x-axis, and it is especially difficult to visibly separate the yellow line (Japan) from the other lines since it is hidden by the dark green line and other lines most of the time. The grid line of 100 is in bold, maybe it was designed to be a reference line, but I doubt if this can be percieved. Besides, we may focus more on the valleys in the year 2020 that stands out, rather than the general decreasing trend in German manufacturing after 2018. *3. Annotation & stand-alone readability.* Neither the x nor y-axis have labels. But we may guess that x-axis represents years, and that y-axis represents manufacturing output by volume (from the subtitle). Besides, there is no clear legend for dark green. We do not know which country dark green represents, although we may guess Germany. *4. Less is more:* The data-ink ratio is low, e.g., the light blue background is not needed since it does not give extra information. The red square in the upper left corner is also not needed. "2015=100" in the subtitle may be confusing and unnecessary since we can see that all the countries output around 100 volumes in 2015. *5. Graphical data integrity & lie factor.* The y-axis is distorted as it does not start from 0, and it is visually deceptive. For example, it seems that for Britain, Japan and Germany, the lowest volumes in the year 2020 are almost only half, compared with the volumes in the year 2015. However, the true decrease is from around 100 to around 80, which is less than perceived.

Part 2. Graphic design

Figure 2

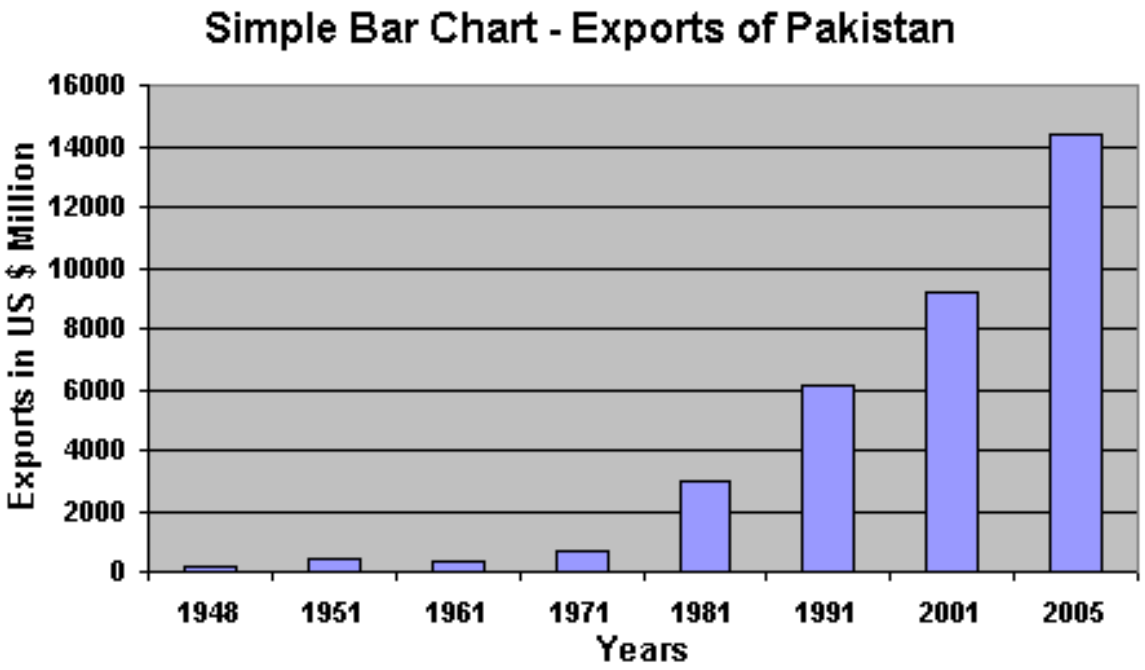
Growth of Orange Trees



Note. The trees are sorted by maximum diameter from smallest to largest. The grey lines represent the average trunk circumferences at different tree ages of the five trees that were measured.

Figure 2 intends to deliver three messages. First, the circumference of an orange tree is positively associated with its age. Second, we can compare the growth speeds of the five different orange trees, tree 3 grows the slowest and tree 4 grows the fastest. Third, we can compare the growth speed of each tree with the average. Trees 1 and 3 are below the average, and trees 2 and 4 are above the average. Tree 5 was below the average at the beginning, but caught up in the end. I used the small multiples instead of a plot with many lines, to make it easy and effective to understand, and reduce the difficulty of comparison for the readers.

1. Keep it simple: Each small plot contains only a black line and a grey line, which is easy to decode and compare the trends. The data points and the grid lines help to anchor the x and y values. The positive association between the circumference and age can be perceived at the first glance. *2. Gestalt principles & visual structure.* The grey line provides a reference and allows for a direct comparison between the tree's growth speed and the average speed. Even the colorblind can identify the difference between black and grey and printing in black and white will not lose information. The trees are ordered by increasing the maximum diameter for easier comparison because the two adjacent plots are more similar. *3. Annotation & stand-alone readability.* Proper X and Y-axis (with units of measurement) titles, facet and legend labels are all clearly provided, as well as the figure title. Thus the figure contains all the main information. *4. Less is more:* The background and colors are designed as simply as possible, and the data-ink ratio is maximized. The grid is non-prominent. The figure does not have unnecessary or redundant information. *5. Graphical data integrity & lie factor.* The X and Y-axis are forced to have the same scale, and Y-axis starts from zero. So we show the true effect size of data in the graph and avoid visual lies. The five small multiples are comparable.



Source: http://zubairacadmy.blogspot.com/2013/12/18-types-of-charts-simple-bar-chart-b.html

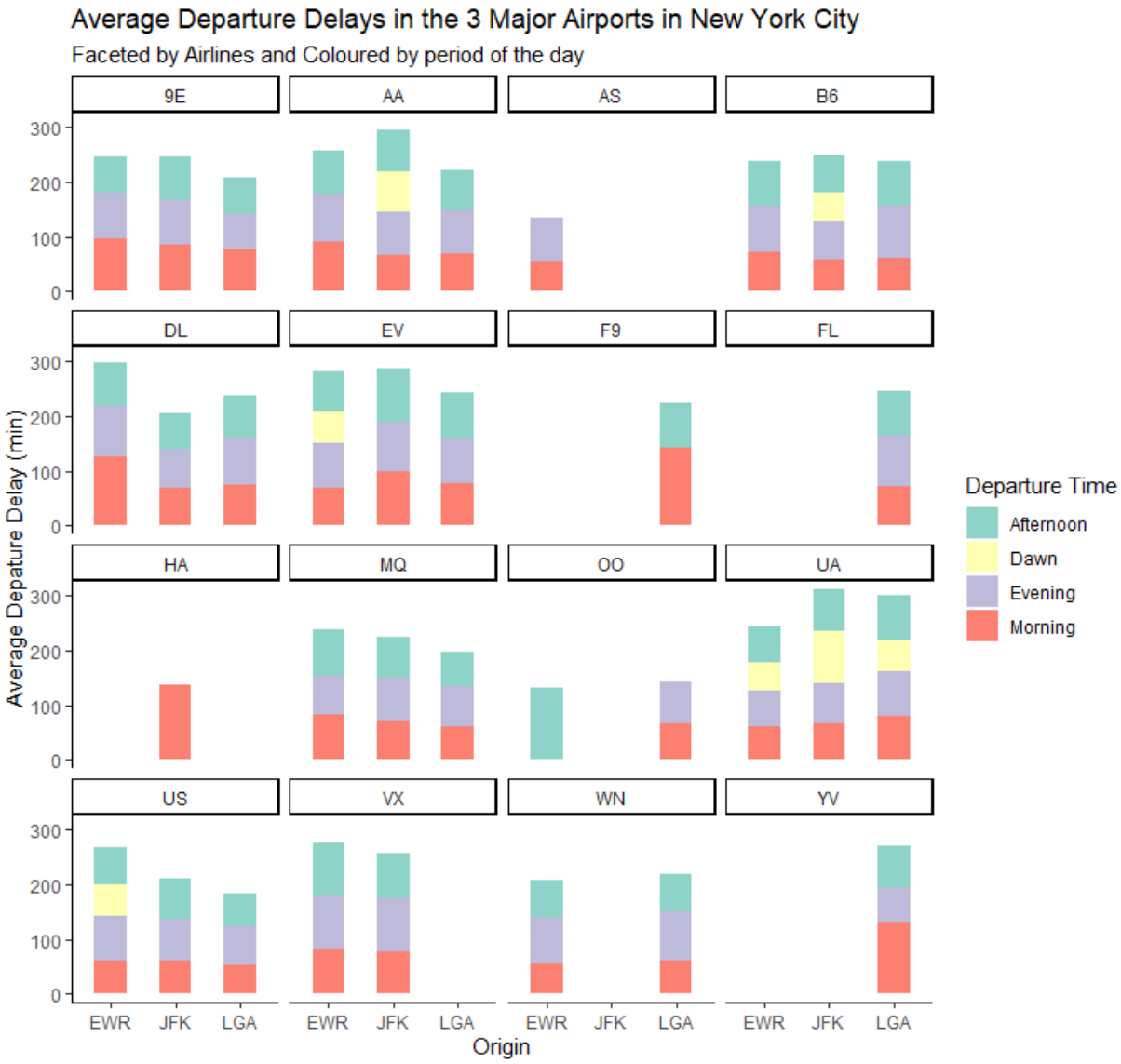
Bar charts are used in expressing the details of a frequency table visually. However, one needs to take a close look at the labels on the vertical and the horizontal axis to understand what the graph intends to communicate. The graph will be critically evaluated based on 5 points. Which include the gestalt principle and visual structure, keeping it simple, less is more, graphical data integrity and lie factor and annotation, and stand-alone readability.

The above graph intends to show the trend in the volume of export from Pakistan from 1948 to 2005. However, there is a lie factor in the scale on the horizontal axis which is intended to show a uniform time gap. It started from 1948, jumped to 1951, and maintained a 10-year interval until 2001 and then a 4-year interval to 2005. Here, the focus of the author is clearly distorted, and this makes the trend inconsistent. Thus, there is a deception when one takes a critical look at the horizontal axis and hence the graphic integrity of the bar chart is flawed.

Also, extra ink is used in colouring the background and showing the gridlines which could have been avoided. The horizontal axis could have been ignored and the individual values represented by the bars could have been annotated on the various bars.

This bar graph is presented in a very simple manner and the values are easy to read even though this could be improved to prevent the audience from struggling to guess the value of bars that fall in between gridlines. The gridlines could have been faded out to serve the same purpose.

Given that this was intended to show a trend, a line graph could have been more appropriate. The axes are clearly labeled, and the title is clearly specified even though it was unnecessary to state that it was a bar graph since it can clearly be seen as such. This to say that the design of the bar graph was good though it could be significantly improved to convey the information it intends to pass.


Average Departure Delays in the 3 Major Airports in New York City
Faceted by Airlines and Coloured by period of the day
9E
AA
AS
B6
DL
EV
F9
FL
HA
MQ
OO
UA
US
VX
WN
YV
Average Depature Delay (min)
0
100
200
300
EWR
JFK
LGA
Origin
Departure Time
Afternoon
Dawn
Evening
Morning
Note: Figures in the table only consider flights which have records of delays more than 20 minutes.
Source: Flights that departed NYC in 2013

The decision on which airline to travel by is crucially dependent on the historical delays of their scheduled departure times. Usually, passengers pay a premium when booking a flight to be compensated for their time spent during the departure delays after a specified amount of time. With this as the background, I decided to check the average delays of airlines that departed the 3 major airports in New York City recorded in 2013 and the time of the day when the delays usually occurred.

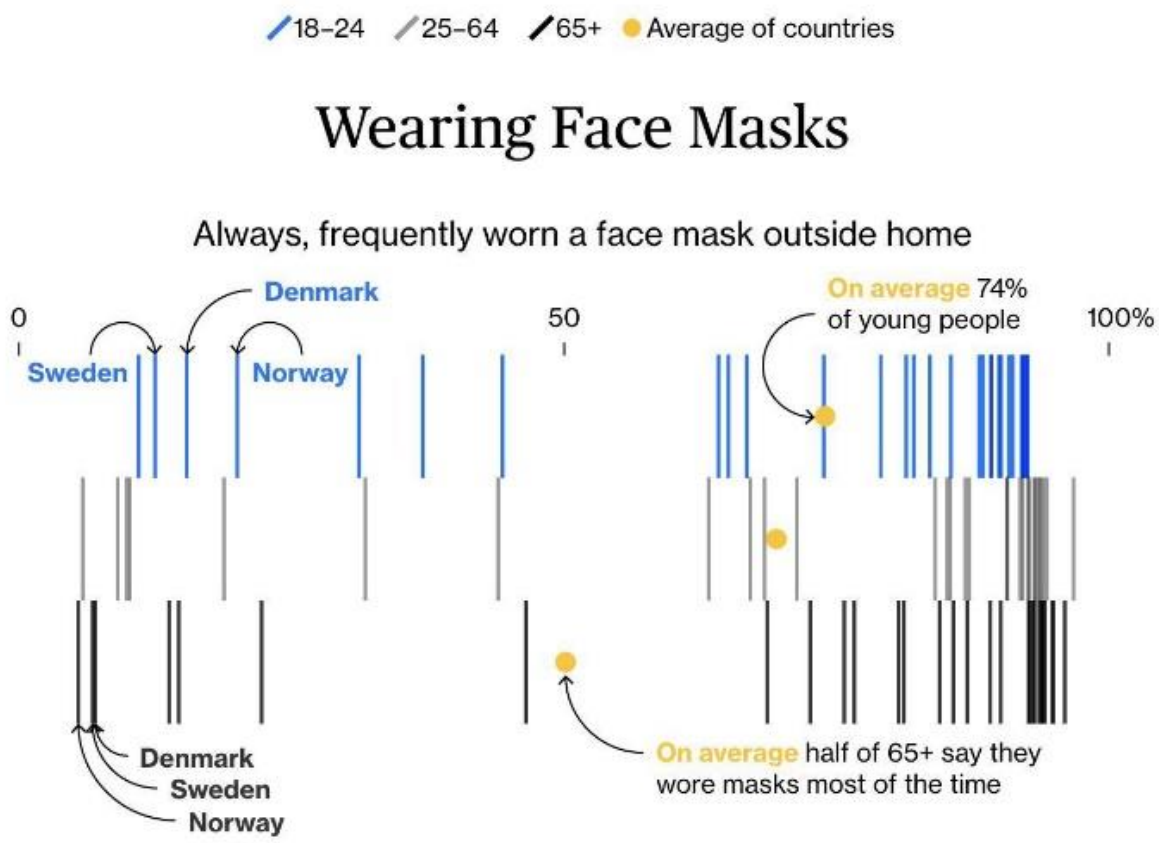
The figure shows component bar graphs that show the proportion of delays according to the time of the day with the average departure delays in minutes on the vertical axis and the departing airport on the horizontal axis. This design was chosen in order to show the count of average departure delays as well as the proportion of delays within the time of the day at the same time.

From the data gathered, it appeared that some airlines departed earlier than scheduled but most people get worried when the flights are delayed. In this regard, the data used included only flights which delayed their departure by at least 20 minutes. This could help passengers in their choice of airline if they care so much about delays at their time of departure. It could also serve as a determinant for insurance companies in deciding how much premium to charge their customers depending on the airlines they choose, the time of their scheduled flight, and which airport they decide to depart from New York City.

This could serve as a guide to passengers in deciding which airline to choose to choose. It could also inform the operator of airlines about the possible decline in their clientele.

Figure 1

Chart showing the adoption of masks across countries



Note: The chart shows adoption of masks across all age groups rising dramatically in many countries over time. From, Bloomberg-Europe edition, by E. He and L. Williams, n.d. (https://www.bloomberg.com/graphics/2020-opinion-coronavirus-young-people-spread-care-about-social-distancing/) Copyright 2022 by Bloomberg L.P.

Reference

Bloomberg-Europe edition. (n.d.). *Young People Care More About Covid Than You Think.* https://www.bloomberg.com/graphics/2020-opinion-coronavirus-young-people-spread-care-about-social-distancing/

The visual structure does appear uncluttered owing to the whitespace. However, it is difficult to understand that the vertical line segments are discrete values denoting a country each mapped along the horizontal axis. The age ranges on the x axis are non-proportional and have been plotted on top of each other without any spaces.

Decoding this chart is a difficult task. While overall the data-ink in this chart is quite balanced, issues of chart junk are present. For example, the countries as discrete values are sometimes overlapping, maybe it would have been better to use a different shape, maybe dots. This overlaying also forms darker lines, so there is uncertainty about how many countries lie therein. Maybe, the graphical distortion could maybe be minimised if a y-axis labelled or even present. Data labels also are overlapping the x-axis which confusingly appears on the top of the chart. The colours for the two age categories 25-64 and 65+ could have been more distinct instead of similar colours. Given the subject of the source article authors probably wanted to highlight the 18-24 age group and have purposely chosen a brighter colour.

While three countries namely Sweden, Denmark and Norway have been labelled, labels are missing for all the other countries represented. However, this chart would still be unreadable if all these countries were labelled. Country labels are also missing for the age category 25-64, while they appear in the other two categories. Category label is also missing in the legend.

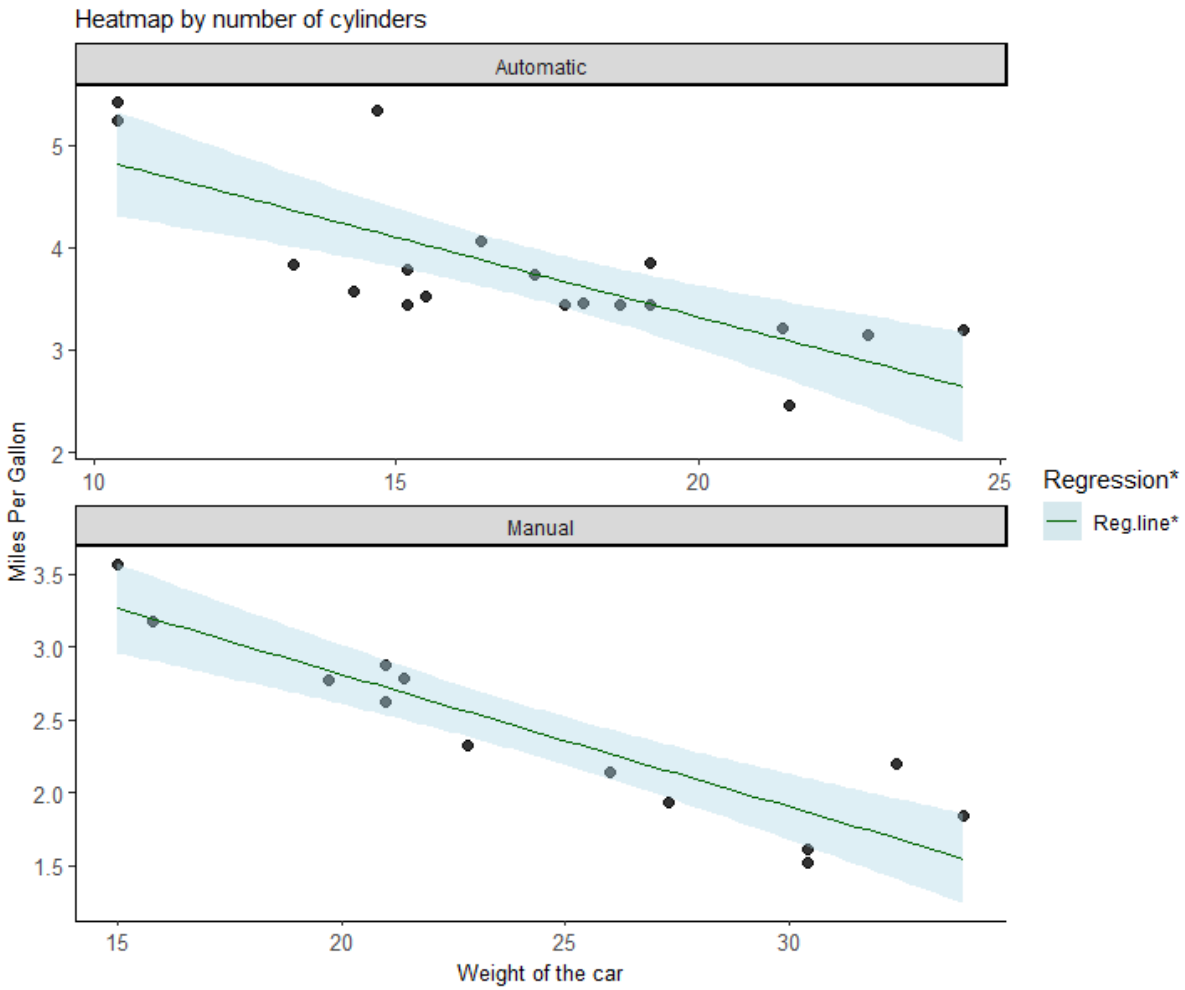
Average response to the question denoted by the yellow dot seems to be an important data point. The labels for these dots are not consistent and could have been removed and articulated better in the legend. However, looking at the 65+ age category, it is unclear if it is actually - ‘averages’, the yellow dots are showing.

The chart title and caption are not descriptive enough to understand the data or the context. The caption ¨*Always, frequently...*¨ adds to this confusion. The legend is not easily found and appears above the chart title. The chart itself is missing source or sample size information.

Part 2. Graphic design

Figure 1.

Relationship between Weight, Transmission and Cylinders on Mileage of a Car



Note: This figure demonstrates the relationship between the Weight of the car, its Transmission type, and number of Cylinders, to its mileage (Miles Per Gallon).

*The regression line shows the relationship between the weight of the car and mileage at 95% confidence interval.

*The data is based on the mtcars default dataset that is provided by Rstudio

Gestalt principles and visual structure

The visual structure of the chart is based on Gestalt principles. The chart comprises 3 variable their relationships can be easily understood by looking at the graph. The is use of whitespace and contrasting colours to make it easy to read and understand.

Decoding and operations

Elements like the direction of relationships is clearly mapped using the regression line, across two different categories. The axes, headers and legend are explicit. There are no overlaps between the points and all the values are clearly visible.

Chart junk and data ink

The data ink ratio is low with only essential elements across all variables in the chart. Maybe the chart headers could have been white background with only the black text. However, the grey background is preserved to highlight these headers. The legend for regression line could have been labelled with the details explained in the notes section. However, doing this would have made the chart space smaller, so the labels had been shortened to accommodate this.

Graphical data integrity and lie factor

Starting the axis of the chart at 0 would make it difficult to highlight the dispersion of the values and their relationship clearly. However, the Axis are labelled and not manipulated deliberately to skew the data.

Annotation and standalone readability

The chart title, subtitle, axis labels, legend, and caption provide all the information required to understand the data context. Relationships can be clearly understood between the different variables and inferences may be made based just looking at the chart.

Portfolio for MAE4000
Component 2: Data visualization
Otávio Mattos

Part 1: Graphic inquisition.

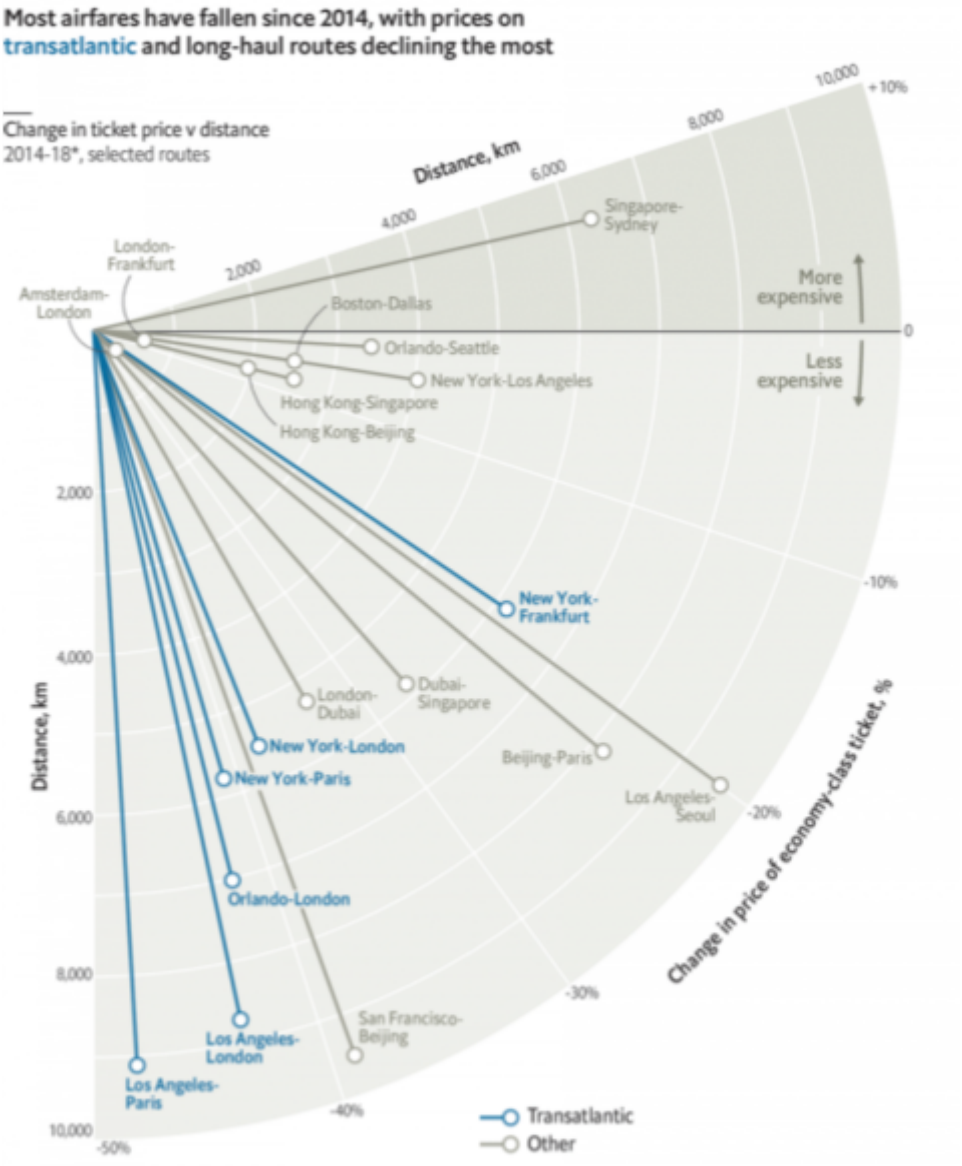


Figure 1. The graph aims to show a relation between distance (in km) and change in price of economy-class tickets (in percentage) ("Why ticket prices", 2018)[1].

[1] Online newspaper article without author, cited according to APA Citation Guide (7th Edition).

The graph in figure 1 was published as an online article from The Economist ("Why ticket prices", 2018). It aims to show a relation between distance (in km) and change in price of economy class flight tickets (in percentage). The graph contains multiple data visualization problems. We summarize some of them below.

First, a key problem is that the graph contains too much visual information to process. For example, we can see 15 lines marking the relation between distance and airfare change (D&FC) for different flights. Different flights, in turn, are indicated through 19 dots and labels. To facilitate the comparison of D&FC across flights, the authors could have used the strategy of "small multiples" (Bock, n.d.) to avoid multiple lines together, or simply used a scatter plot with a regression line where each dot stands for a flight.

The graph also brings additional comparisons, making it overly "busy". The lines were distinguished by color, i.e. blue for "transatlantic" and dark gray for "other", suggesting that we should also compare D&FC between these categories. A line graph or a scatter plot with one regression line for each category could have facilitated cross-category comparisons.

We also think that the "temporal component" of the data could have been implemented differently. Instead of having the variable "airfare change" with percentages, they could have shown different scatter plots with the relation between distance and airfare *per year* (i.e., a scatterplot for each year, from 2014 to 2018). In this way, we would have been able to see how airfare has been falling throughout the years.

Last, but not least, the relation between distance and airfare change in their graph requires us to consider two visual dimensions: the *direction* of the lines (to indicate fare change), and *length* (to indicate flight distance). A standard scatter or line plot would have just required direction, making the "decoding" of the correlation simpler.

Part 2. Graphic design

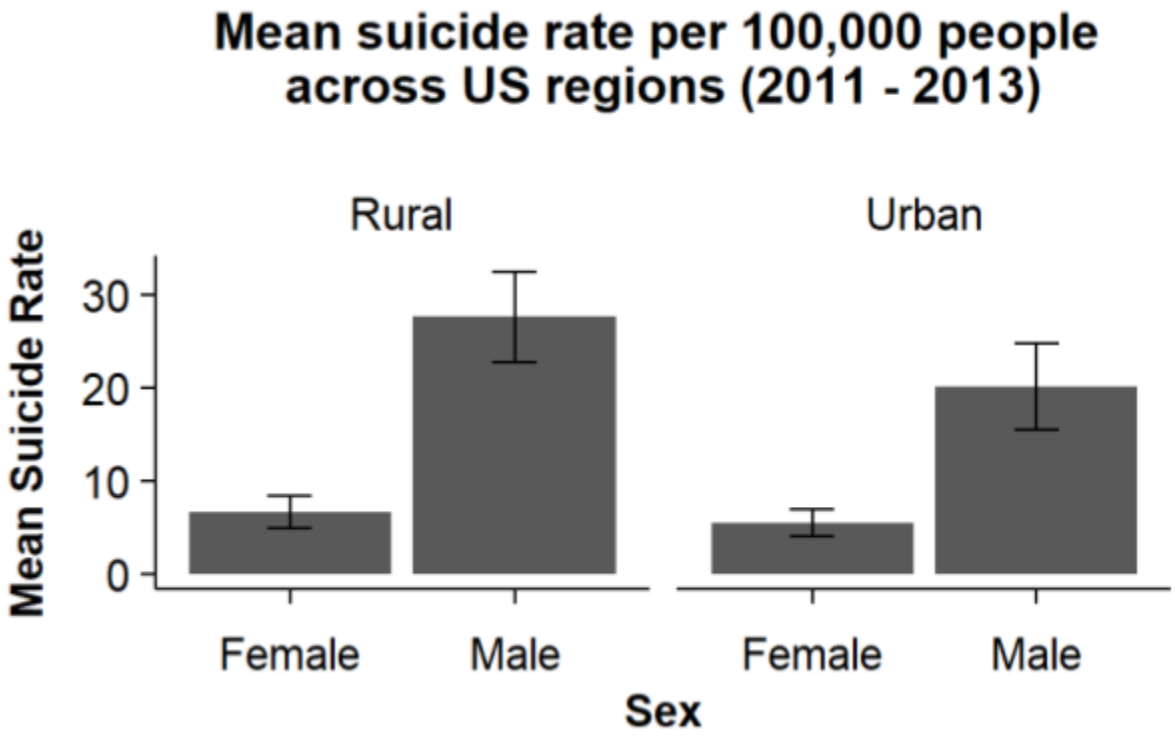


Figure 2. Mean suicide rate per 100,000 people across 10 HHS regions in the United States (HHS: Department of Health and Human Services), between 2011 and 2013. Suicide rate distinguished by sex and rural-urban status. Error bars stand for standard deviation. Data publicly accessible in R under the name of "USRegionalMortality".

Figure 1 intends to show the mean suicide rate per 100,000 people across HHS regions (Department of Health and Human Services) in the US. The data was collected between 2011 and 2013. The bar graph involves three variables: sex, urban-rural status and mean suicide rate. It shows that suicides among men are, on average, higher than among women across HHS regions. The graph also seems to indicate that the suicide rate among men in rural areas is slightly higher than among men in urban areas (note, however, that the standard deviation for rural and urban men partially overlaps; not all rural regions regions have higher male suicide rates than urban regions).

Since our goal was to make cross-category comparisons (sex and urban-rural status), we decided to use bar graphs — ruling out graph options that are especially tailored to explore relations between different numerical variables (e.g., mean suicide rate and time). We could have used box plots instead of bar graphs. Box plots would have allowed us to show the dispersion of regional suicide rates from the median. However, we opted to use bar graphs for the sake of simplicity: we wanted our graph to convey the message that, overall, individuals from some categories are more prone to commit suicide than others, rather than reflect on regional variability within categories.

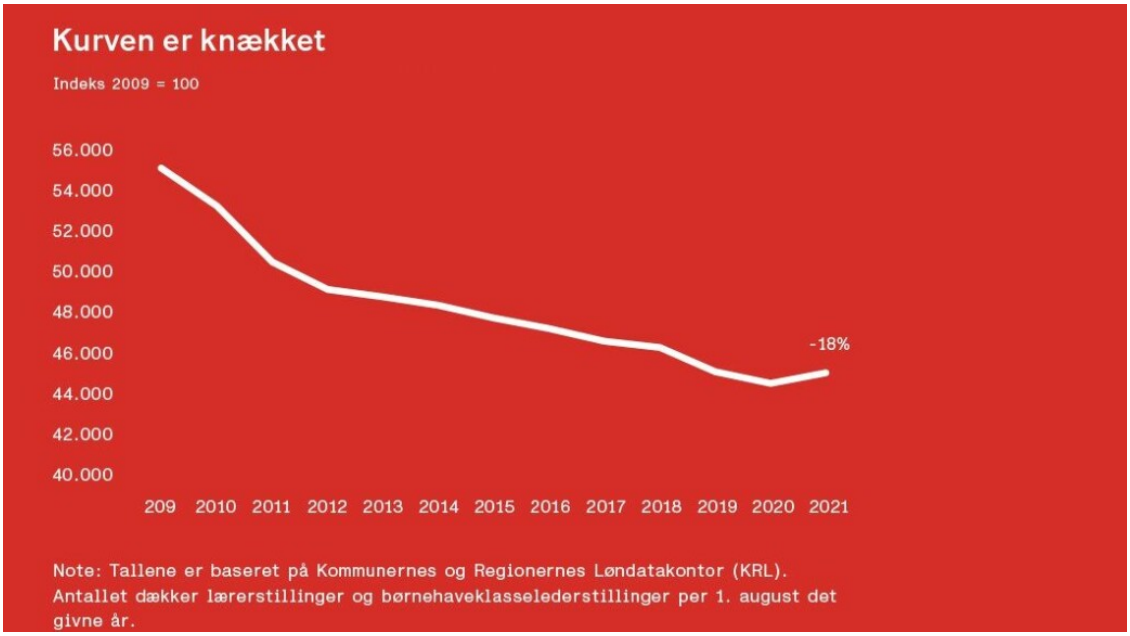
Finally, we decided to have a black and white graph since color distinctions would be informatively redundant with the graph labels. Finally, we decided to distinguish statuses through panels and sex through bars, instead of the opposite. The main suicide rate difference is between men and women and having sex distinguished through bars highlights this distinction —after all, we show it two times, one for each urban-rural status.

Data science exam portflio component II data visualization

Student-ID: 110560

1 Bad graph

Figure 1: *Bad graph*



Source: https://www.folkeskolen.dk/finanslov-folkeskolen-nr-15-2022-kommunal-okonomi/antallet-af-laerere-i-folkeskolen-er-steget/4675003

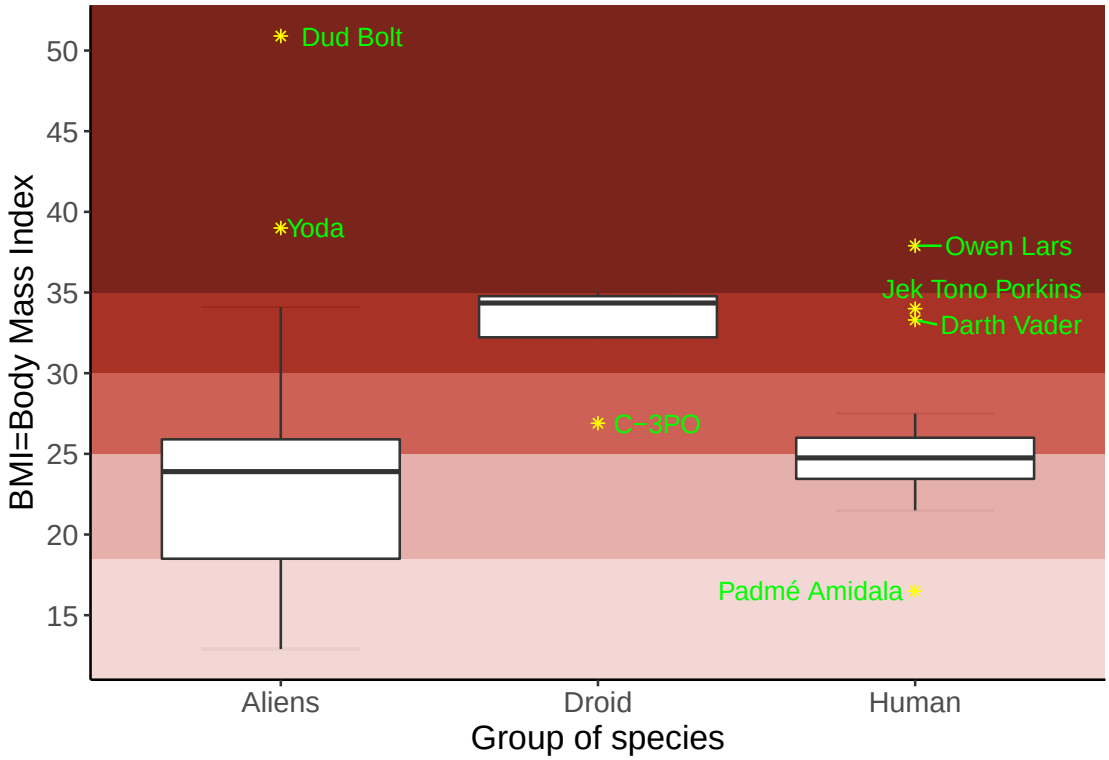
First some context for figure 1 for why this graph have been maded in the first place. The graph has been maded to underlined a pointe of fewer people employed as teachers in the public schools. Hence it should lower the quality of the school system. The title of figure 1 is that "The curve has been broken", and implying that more people are employed as teachers. So the idea and the graph itself is quite simple, but the graph in figure 1, is a bad graph in multiply ways. Obviously the graph itself with all the flaws, but also in what is should represent data and conclusion about. The only good thing that figure @ref(fig_badgraph) has, is it simple and quite easy to see what they inteded to do.

The flaws is many in figure 1, and starting with the x-axis, which should be the year of 2009 to 2021. But the first observation is in 209, which should have been 2009. The note in figure 1 says that the data comes from register of salaries, so the years should be correct. The y-axis is started at 40.000 and ends with 56.000. There is no unit of the y-axis, and the title of the graph do not indicate what the units are. The choosen range of the y-axis makes the decline looks bigger, than if the range have been at absolute minimum at 0 to 56.000. By the choice of range, make the figure 1 a lie-factor of $\frac{(44.000-55.500)}{(0-55.500)} = 20.91\%$. Furthermore the label of the y-axis says "Indeks 2009 = 100", which is a way to scale numbers to each other, if there was multiply variables. But since the y-axis is way past 100, and 2009 do not exist at the x-axis, the index become meaningless. This label is just adding to confusing and understanding. At the end of the curve there is a label with the number "-18%". This number do not reference to anything or given an explanation. You have to guess that it should be the percentage change since 2009. This values could have been seen if it the graph have been the index as label of the y-axis suggest. At the right side of the graph there is a huge gap, where the graph is not utilized, and then it is just a waste of space.

When looking at the information that the graph should provided, the development does not make sense if it is not related other things. It would omit any confounding variabel, since it only shows the number of teachers in figure 1. In this case the confounding variable could be number of pupils or classes, and then the teacher could be compare to the development in pupils and/or the number of classes. If the quality of teaching should have gone down the ratio between $\frac{teachers}{pupils\ or\ classes}$ should go down. So the curve of the index for teachers should have been compared to an index of pupils or an index of classes. So with these curves the level of information would have grown. The figure 1 is a figure that cannot stand alone to drawn the suggested conclusion.

2 Good graph

Figure 2: *Obesity among Star Wars characters*



To answer the second part of the visualization component the graph in figure 2 have been constructed. The graph investigate the obesity of characters in the Star Wars universe. It investigate whom of the characters that have a high *Body Mass Index*(**BMI**), how the distribution of observations within each species is, and all this is compared to the thresholds of obesity defined by the human species. All data is from the *Star Wars* dataset build into the *tidyverse* package in R. Figure 2 has been build by constructing the **BMI** by the equation:

$$**BMI** = \frac{weight\ in\ kilo}{(height\ in\ meters)^2}$$

and the group of species are build to compare aliens, droids and humans with each other. In the star wars dataset the groups of droids and humans are defined as unique species, while the alien group is rest of the species. Within that category there are different species, such as Wookies, Ewok, Gungan, etc. All types of aliens are put together since the number each species are too small to form its own group.

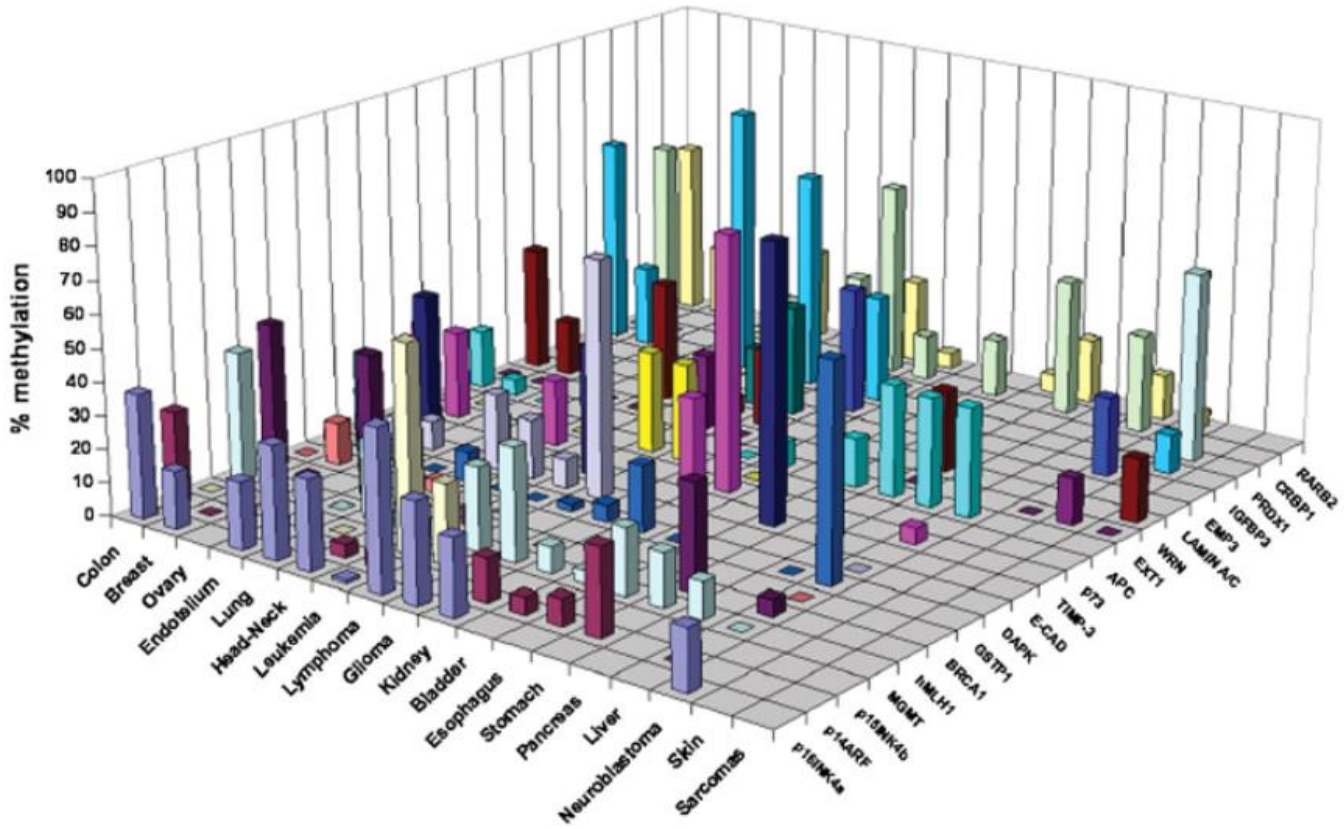
The **BMI** is chosen as an indicator of obesity, since it takes the characters' height into account when comparing the weight. To give the best comparisons of **BMI** across groups of species a boxplot have been chosen, since it gives the distribution of the group, and only one observation is removed from the distribution, since it was an extreme value. The observation was *Jabba the Hutt* with a **BMI** at 443.4, which would have distorted the boxplot too much, so it was better to remove it. The distributions are based at 31, 4, 22 observations in the group of aliens, group of droids, and group of humans. The box plot also indicates what is the normal range of **BMI** within each group of species.

The severity of obesity are related to the human scale of obesity, and the different categories of obesity are painted a shade of red, with a darker color as indicator for severe obesity. A **BMI** at or higher than 25 indicates overweight and **BMI** at 30 or higher indicates obesity.

In figure 2 droids are generally obese, but mostly because they are made of metal. Humans are mostly in the range of normalweight and overweight, with some obese characters, and Padmé as underweight. Within the group of aliens most of them are normalweight, but interestingly are Yoda obese.

Figure 1

A CpG Island Hypermethylation Profile of Human Cancer



Note. This graph shows a profile of CpG island hypermethylation in human primary tumors. The Y-axis shows the frequency of hypermethylation for each gene in each primary. Graph adapted from Esteller, M. (2007, April 15). Epigenetic gene silencing in cancer: the DNA hypermethylome. *Human Molecular Genetics*, *16*(R1), R50–R59. https://doi.org/10.1093/hmg/ddm018

Part 1. Graphic Inquisition

1. Gestalt principles & visual structure: People tend to dislike pie charts and three-dimensional figures. It makes decoding difficult and makes it complicated to read off values. This 3D-structure makes it hard to distinguish the percentage of each bar. Some bars are hidden behind others. This graph, on one hand, might have enabled to show more data at once, but makes the comprehension more difficult on the other hand.

2. Decoding & operations: For me the gridlines aren't helping to easily read off the values. It's making the graph look busier. For decoding purposes, this isn't handy.

3. Chartjunk & data-ink ratio: There is quite a lot of text on the graph, which is distracting. The colours of the bars are probably supposed to differentiate between the bars, but are actually distracting and standing out.

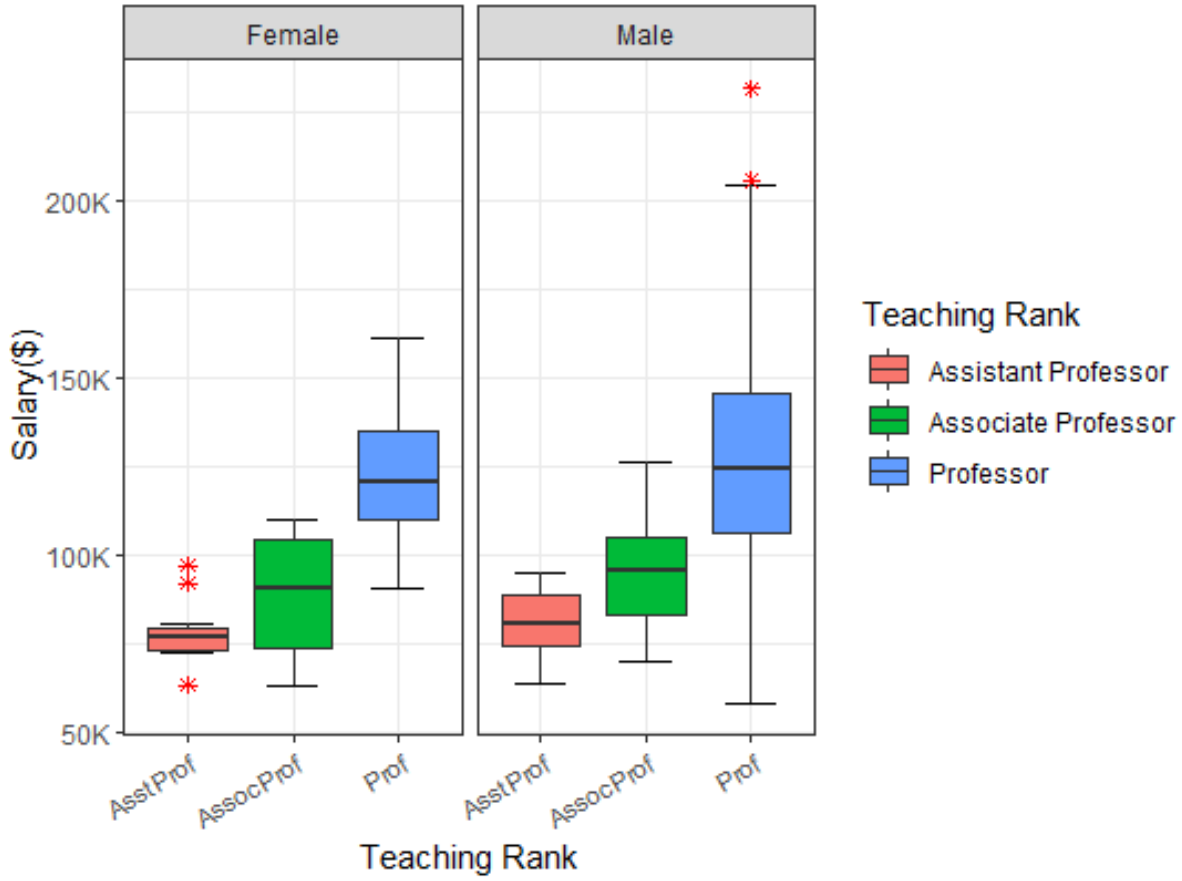
4. Graphical data integrity & lie factor: The chosen format and design might not be the best way to represent the data. There is too much information in one frame. The labelling isn't very clear. What do the unnamed axis represent? It's hard to differentiate between the values of the (hidden) bars, it isn't keeping the data intact.

5. Annotation & stand-alone readability: The annotation is incomplete, only the y-axis is labelled. There is no legend to explain what the colour of the bar stands for. This figure also scores low on stand-alone readability. It's unclear what the graph represents, understanding it requires some text references. There are also many abbreviations (e.g. EXT1, E-CAD, …) .

Part 2. Graphic design

Figure 2

Academic Salaries (2008-09) of Teaching Staff Classified by Gender



Note. 'Salary' is an aggregation of the monthly salary (in dollars) over a period of nine months (2008-2009). The teaching rank classifies the academic staff of a college in the U.S. in three categories: 'Assistant Professor', 'Associate Professor' and 'Professor'.

This graph displays the salary distribution for the female and male academic staff of a college in the U.S. I opted for a box plot graph, because it gives a quick overview of the dataset. Also, because it makes comparing distribution between two groups (e.g. male-female) easier.

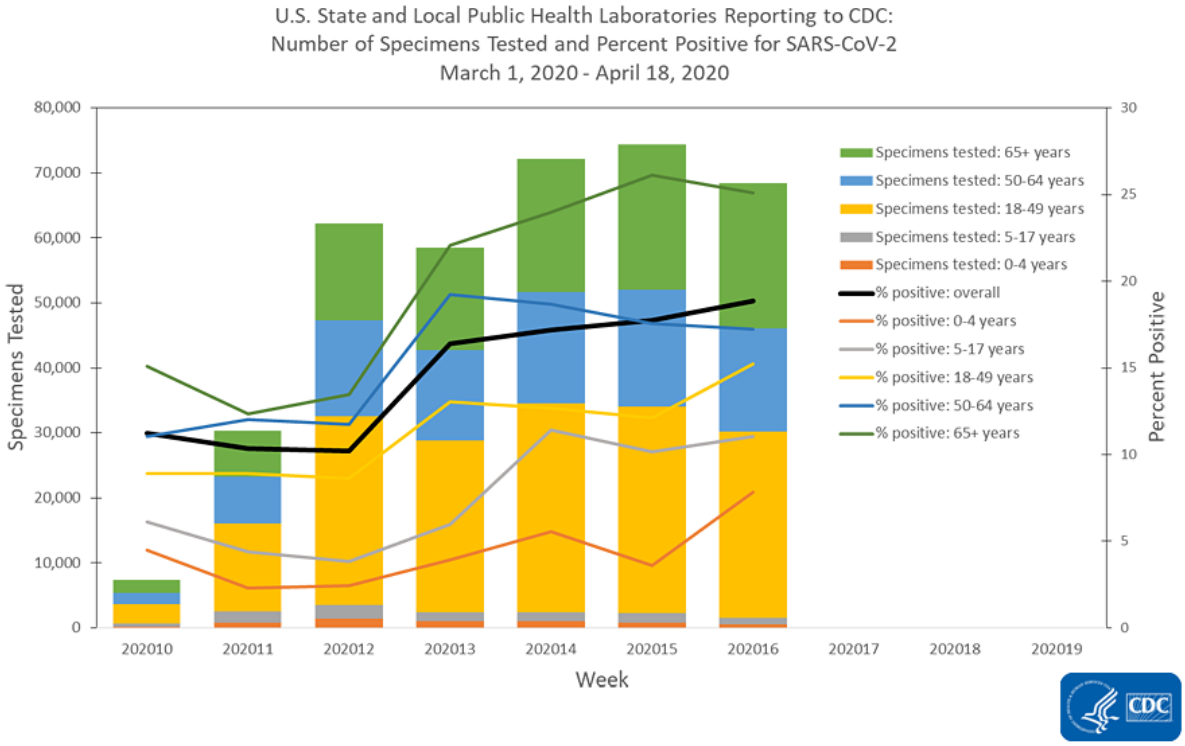
The axis have the same scale for both facets to make it easier to read off values. For readability purposes, I decided to rename the ticks of the y-axis: 100K is easier and faster to read than 100000. For visual purposes, I decided to make a distinction between the three different teaching ranks by using colours for the boxplots.

The red asterisks represent outliers. Not all the outliers were included in this graph, just so that the proportion of the boxplot in the graph wouldn't shrink. I signalled the outliers, but didn't add their value to avoid creating Chartjunk. The numbers would be distracting and wouldn't really create meaning. It's the overall picture we're looking at, not specifically individual cases.

The graph shows that the median for the male teaching staff is higher than the female teaching staff for all three ranks. What pops up is that there is a wider distribution for the male professors' salaries. This means that there is a big difference between the minimum and maximum salary for this particular group (e.g. male professors).
If we compare the teaching staff among themselves, meaning in the same group or facet, we notice that there is an observable difference between their salaries. The median for a professor's salary is respectively higher than the median for an associate professor's and assistant professor's salary.

Figure 2

Public Health Laboratories



Note. "CovidView Summary ending on April 18, 2020". From "Centers for Disease, control and prevention", 2020.

Sources:

Centers for Disease Control and Prevention. "COVIDView Summary ending on April 18, 2020", From: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/past-reports/04242020.html

Part 2- Critique of graph

The chosen graph intends to show the number of Specimens tested and the percent positive for SARS-CoV-2 with the time interval of March 1, 2020- April 18, 2020.

When it comes to being able to read the graph without any background information, after reading the header and labels it seemed like a graph that would be relatively simple in a sense. But after having a closer look the x-axis showing weeks respectively should not be a problem, but it becomes very unclear as the different blocks are not divided into months and years, such as “202010”->”2020/10”. This would make it clearer to the reader, and the overall impression would increase.

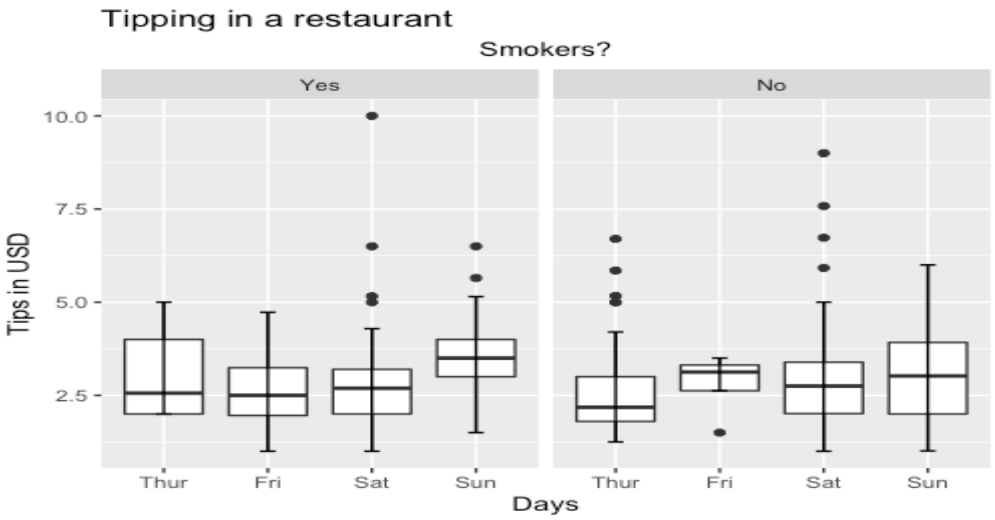
For the Y-axis(es), first, there are two, one shows specimens tested, which seems to be the Y-axis the graph sticks to, on the other hand, there is also another Y-axis that shows the percentage of positive tests. In a way, this is somewhat forgotten by the fact that the factors are inside the graph itself and take away a bit of the purpose.

The graph itself that has been used contains both stacked plots and lines. To start with stacked plots, these make it very difficult to compare with the different weeks, since (ex: low numbers are hidden in the bar plot itself. On the other side, the stroke of genius is to choose the same colors for the line and bar plot, it is understandable that this should show similarity, and which belong together- which it does to some degree, but on the other side, the same colors go into another, making it hard to follow.

A graph should have positions on various factors so that the content is systematically easy to see. It is important to think of visualization as storytelling, which should simplify statistics, not make it harder to understand compared to seeing it in a table.

Figure 1

Average tip in a restaurant



Note. Own graph, STUDID:97523

Part 2. Graphic design

The first clear intention I had when making the graph was to make it look as easy as possible but show clearly what the intention was. I first tried to make a point graph, but the observation was too many and difficult to read because of the overlapping points. I then decided to go for a boxplot.

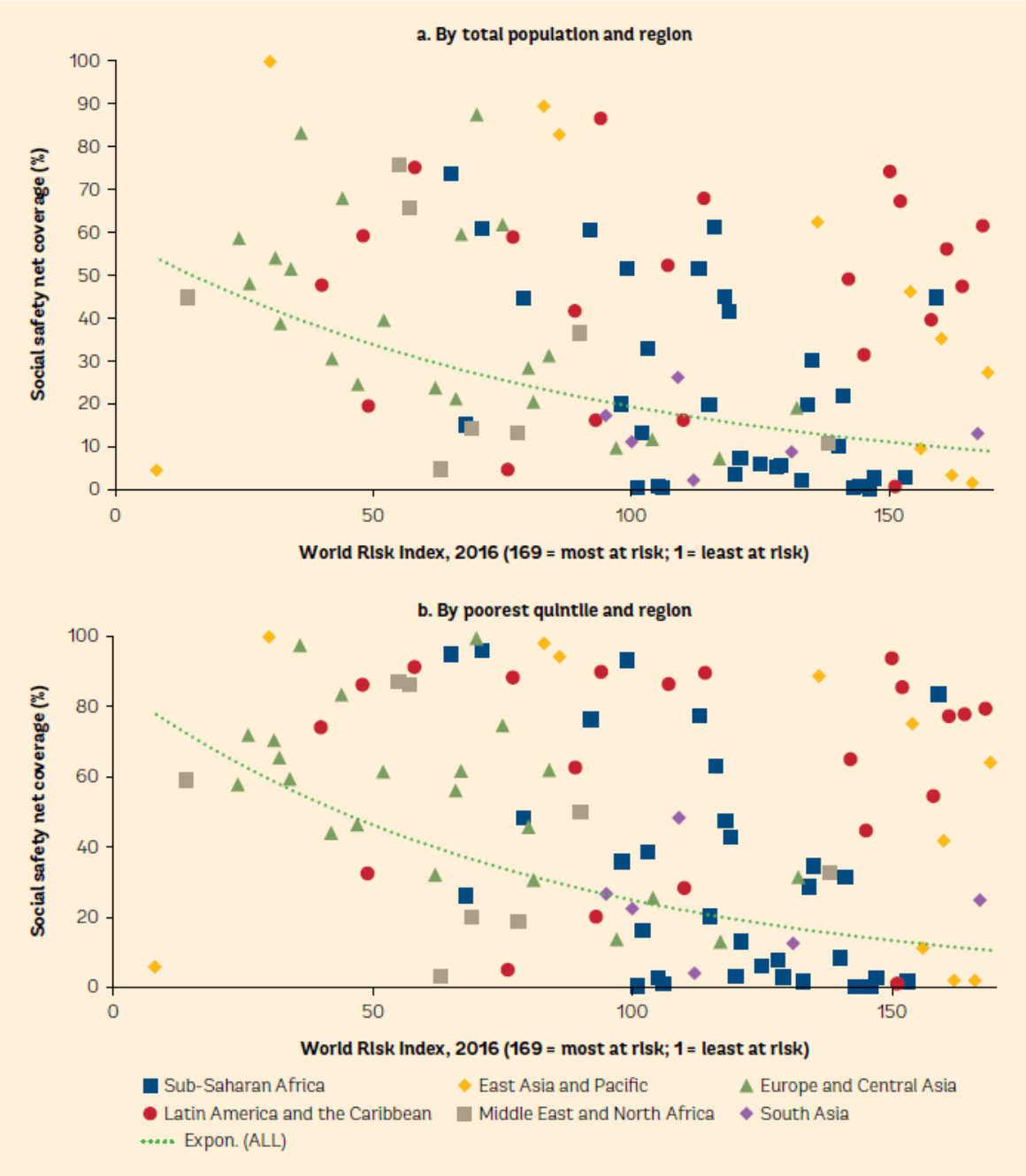
Background information about the graph: A waitress recorded tips and bills from the restaurant she worked at, the chart made is intended to show the difference between tips when it comes to days of the week, and if there were smokers in the group.

The choice of using the factor with smokers to divide the boxplot was to see the difference in tipping according to when it is split between if there were smokers in the group or not.

The graph also intends to address on which days the average tipping is highest, as we can see from the graph, the highest average of tipping is indeed Saturday and Sunday when it comes to the group with smokers, on the other hand- the group of non-smokers shows that Friday and Sunday have the highest average, with most outliers (highest tippers), Thursday and Saturday

The choices of design are as stated earlier, less is more. Overdoing colors, functions, etc. can often result in the message and the meaning disappearing and shifting focus. The stand-alone readability is strong because of the title and labs, that in some sense “describes” what the graph intends to show. This also comes with clear decoding and operations. The annotations in the graph are clear from the "xlab" and "ylab". It can be discussed whether the graph contains valuable information- but after trying out different themes, and trying to do it in different ways, this was the most understandable way, and where the main purpose came across in a clear way with minimizing chart junk and having a theme in mind “The simpler the better”.

FIGURE 5.3 Ranking of Natural Disasters and Safety Net Coverage



Sources: Garschagen et al. 2016; and ASPIRE database.
Note: Social safety net coverage is based on the latest year for the ASPIRE database (all programs). ASPIRE = Atlas of Social Protection: Indicators of Resilience and Equity.

From:
Copley, A. (2019, April 12). *Figures of the week: The state of social safety nets in Africa.* Africa in Focus. https://www.brookings.edu/blog/africa-in-focus/2018/04/12/figures-of-the-week-the-sate-of-social-safety-nets-in-africa/

Component II: Data Visualization

A. Graphic Inquisition

Gestalt principles and visual structure:

The first impression looking at the graphic in page 3 is that everything is messy and haphazard, almost like confetti thrown over a rectangular space. All the elements are vying for attention. There is no discernable pattern. There is no sense of similarity or grouping even though the color and shape choices attempt to convey it. There is no symmetry.

Keep it simple: Decoding and Operations

Very difficult to crack the meaning behind the graph. The creators included a legend and made color and shape distinctions to represent the regions, however there were some over-lap in the shape, for example the diamond and squares. The color green representing Europe and Central Asia region was also used for the exponential curve. The scales on the y-axis in the two graphs are different. Very complicated to decode and understand.

Tufte Concepts: Less is more Chartjunk & Data-ink ratio, Graphical data integrity, Lie-factor

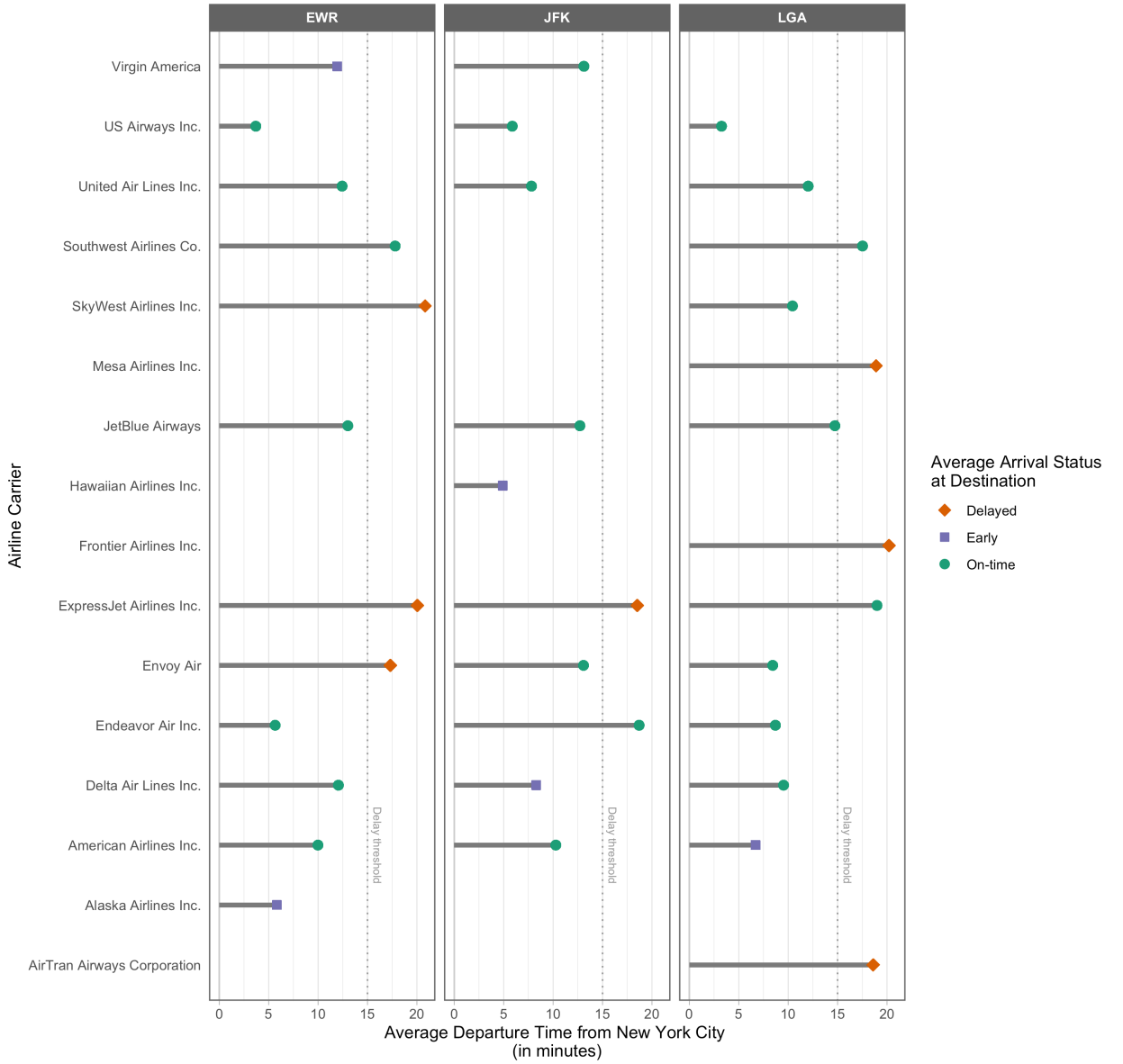
The graph is overplotted and could have benefitted from using small multiples. Better yet, the message "The higher the world risk index, the lower the social safety net coverage" could have been delivered with less chart-junk. Simpler background (without the yellow color) is better as well. Color-choices may not be color-blind friendly. It's not clear whether each data point represents a country in the region. Might have issues with printing. The y-scales are different, making the graph below larger compared with the one above.

Annotation:

The graphic is labeled and annotated. The source was cited. However, doesn't follow APA style. It takes a while to understand the title of the graphs in relation to what is being plotted vis a vis World Risk Index. Poor stand-alone readability.

Figure

Airline Performance: Departure Time vs Arrival Status



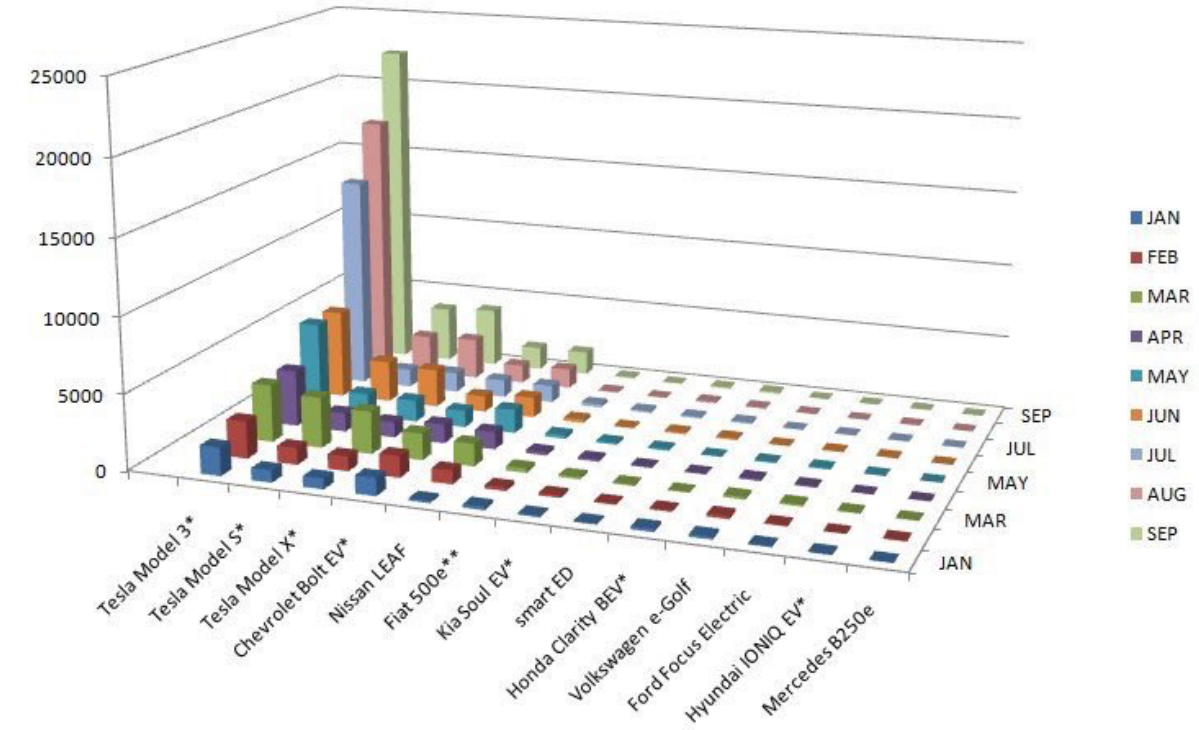
Note: Figure demonstrates how mean departure status from New York City airports affects airline performance as mean arrival status at destination. Delayed status > 15 minutes as defined by Federal Aviation Administration. EWR = Newark Liberty International Airport, JFK = John F. Kennedy International Airport, LGA = LaGuardia Airport, NYC = New York City. Data Source: Wickham H (2022). *Nycflights13: Flights that Departed NYC in 2013.* R package version 1.0.2, (https://github.com/hadley/nycflights13).

B. Graphic Design

The design concept of my graphic in page 4 stems from the following questions: “Which airline departing from New York City, will bring you to your destination on time?” and “Will departure status affect arrival time?” I used flights and airplanes from nycflight13 package. I added a few variables in the data set including the mean departure/arrival delays as well as a status based on the Federal Aviation Administration definition of 15 minutes. By doing so, I could minimize overplotting of data and be able to send the info I wanted to convey – airline performance vis a vis delay in departure and arrival. For design choices, I tried to incorporate the principles we learned in class. Symmetry, simplicity and grouping. I used a light background theme and removed the horizontal grids. I kept the vertical grids because I felt it is useful but I made sure to use a lighter color so as not to cause distraction. By looking at the graph, one can tell which airline performs better depending on which airport it was departing from. Most airlines that depart on time, arrive to their destination on-time, and vice-versa. However, some airlines, even departing late, will try to make up the delay and arrive at the destination on time. And some (i.e. Alaskan Airlines) on average arrive early.

Annotation: Since I meant this figure to be read by lay-people, I used casual language (i.e. Average) rather than Mean. I also used abbreviations for the three airports which I included in the notes.

Color /shape choices: I used a color scheme in this case to highlight the arrival status (qualitative data). I made sure to choose from the color-blind friendly palette of R brewer. Also, I used different geom-point values, so in case this was printed in B/W, the difference can still be seen.



Source: https://twitter.com/Austen/status/1054094811542708225

1. Gestalt principles & visual structure:

   Unnecessary 3D bars and the wrong graphic type. They could have easily showed the data with a line graph, instead of a 3D bar graphic.

2. Keep it simple: Decoding & Operations:

   Unnecessary 3D bars in the 3D background and many colors are making the graph hard to understand. Grid lines are helping to read only a few taller bars in the graph.

3. Less is more: Chartjunk & data-ink ratio:

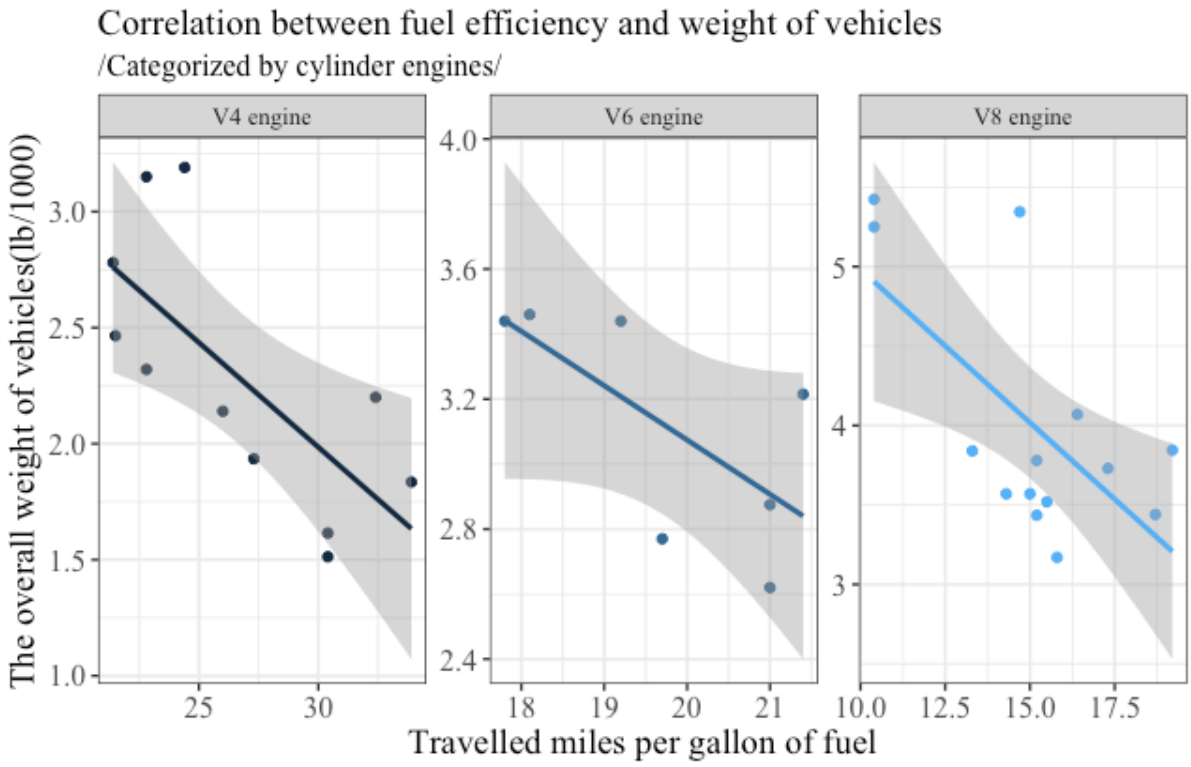
   Half of the 3D bars/car models are in same size and there is no clear difference between them. If these car models were placed under a single variable such as "Other", the reader would immediately notice the type of car they wanted to highlight.

4. Graphical data integrity & lie factor:

   Scale on the vertical axis is fine, but what the numbers are telling is unclear. On the 3rd-dimensional axis, months are messy. It is not clear why even months were skipped. Again, a lot of “flat squares” look like having 0 values on the vertical axis in the graph.

5. Annotation & stand-alone readability:

   Months on the 3rd-dimensional axis and legend that shows which color is representing which month are overlapped. There is no title. Readers only can assume that graph is telling about cars. What the numbers on the vertical axis are telling is unclear. It might be the number of cars has been sold in these months or the total sales of different models of cars. Who knows. On the horizontal axis, what the single and double stars after the cars' names telling is also unclear. And there is no information about the data from which year. Readers can assume the year only from the source, which is tweeted year on Twitter, not from the graph itself.


Correlation between fuel efficiency and weight of vehicles
/Categorized by cylinder engines/
V4 engine
V6 engine
V8 engine
The overall weight of vehicles(lb/1000)
Travelled miles per gallon of fuel
1.0
1.5
2.0
2.5
3.0
2.4
2.8
3.2
3.6
4.0
3
4
5
25
30
18
19
20
21
10.0
12.5
15.0
17.5

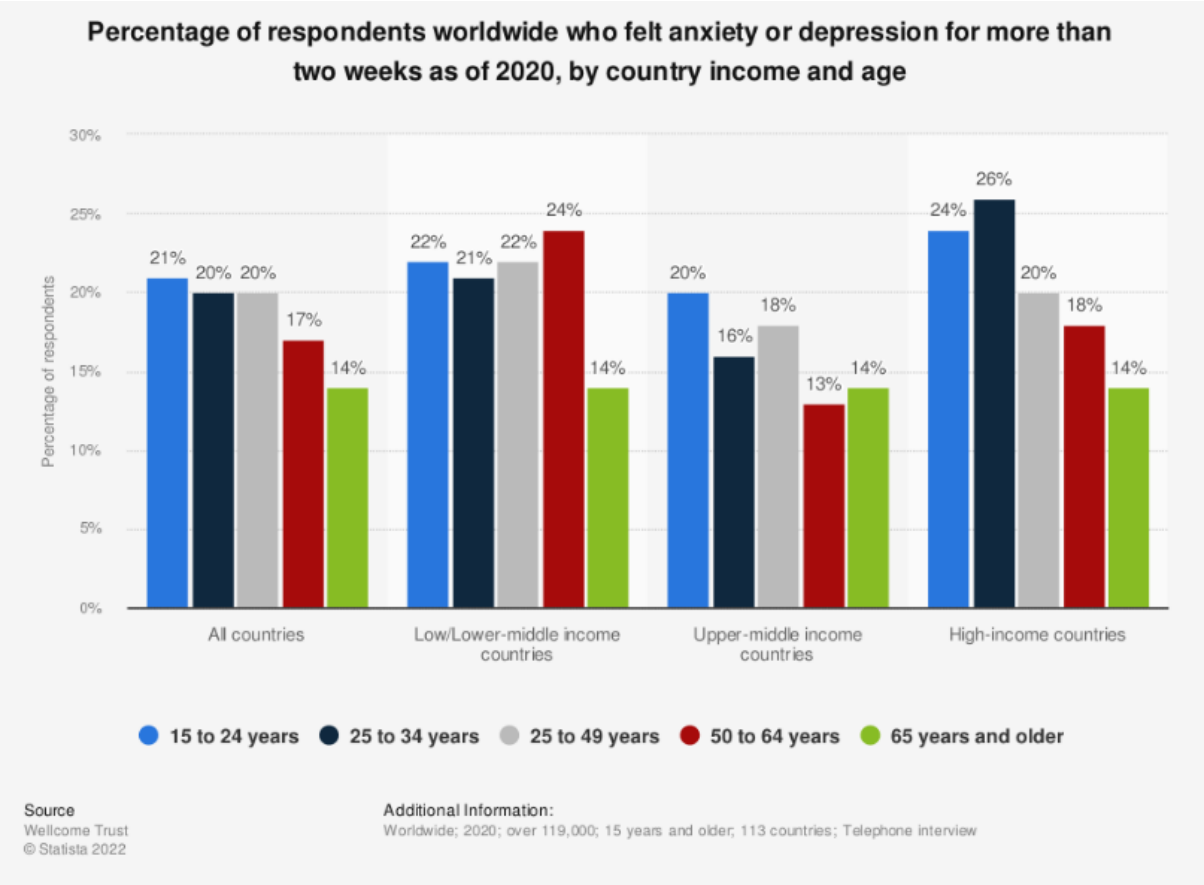
The graph is showing the correlation between fuel efficiency and the overall weight of cars by the number of cylinder engines. There are 3 different types of cylinder engines which are V4, V6, and V8. The trend line or regression line is showing that if the overall weight of cars is lower traveled miles per gallon of fuel increases. This trend is the same across the different numbers of cylinder engines of cars. However, from the deviation of the line, we can tell that correlation between fuel efficiency and the overall weight of cars is less within the vehicles that have 6-cylinder engines.

I used geom_smooth to show the trend line and geom_point to show each represented car. I have tried not using scale = "free" to show the exact deviation of the trend line, but it was making the graph ugly. Thankfully, using scale = "free" did not really affect showing deviation of the trend line within the types of cylinder engines. Since I used scale = "free", I wanted to keep the squared grids on the background to make it easy to see the values of both axes.

Part 1. Graphic Inquisition

Figure 1

Percentage of respondents worldwide who felt anxiety or depression for more than two weeks as of 2020, by country income and age



Note. From (https://www.weforum.org/agenda/2022/10/what-is-world-mental-health-day/).

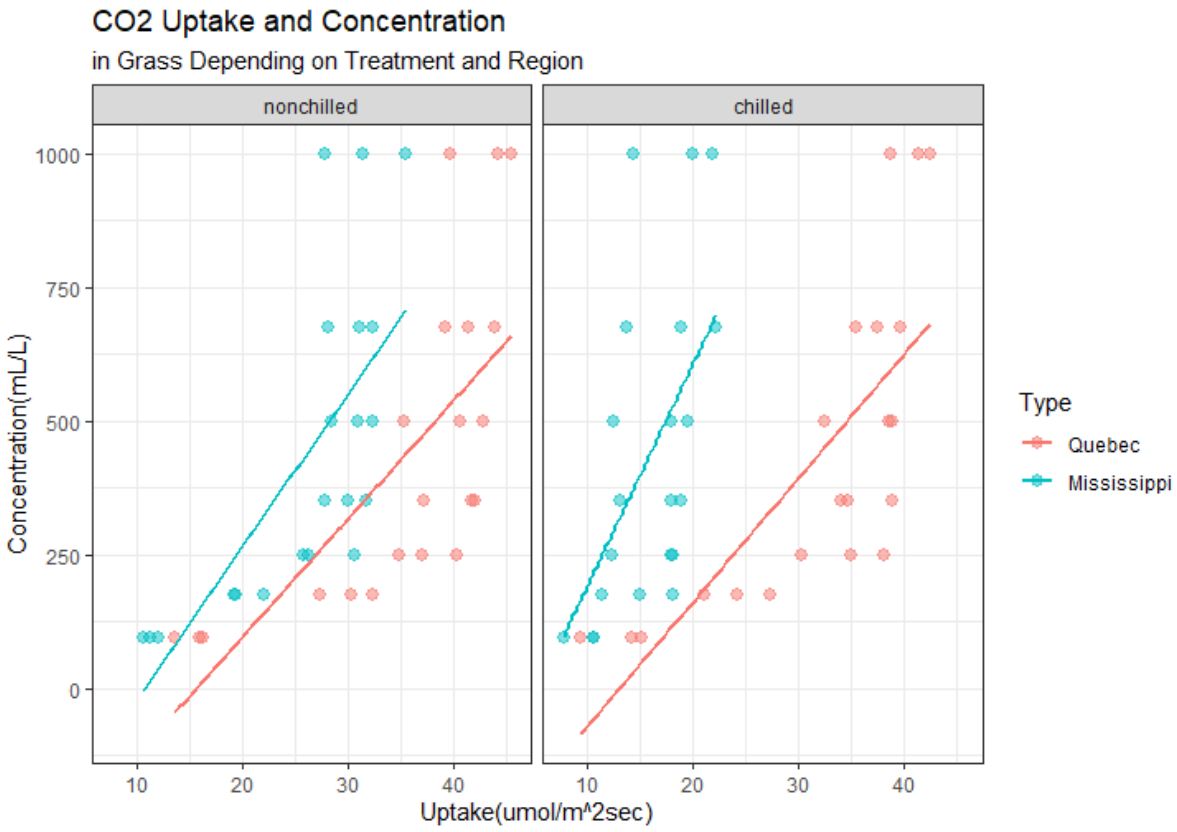
Firstly, this graphic has merit in terms of Gestalt principles and visual structure. Close 5 bars show the proximity principle which helps readers are able to group those in the same income category. It is also related to the similarity due to every bar has the same rectangular shape that helps to group and recognize the bars. The fact that this figure only consists of bars also earns simplicity. This figure achieves the “keep it simple” principle as well. Different lengths of bars indexing percentages of respondents are effective for numerical representation. Upright directions of bars also ease decoding which shows its increasing and plus value. Besides, grey and dim grid lines give 5 percent points of reference of anchor for readers to scan values. It is also simple to see that bars are projected horizontally without any overlap. Distinguishable 5 colors are proper enough to show different age groups. In line with that simplicity, the Data-Ink ratio is quite good. It is likely that there is no redundant or non-data-ink. There are also no chart junks not related to the core message: percentage of respondents by income and age.

However, this figure has some flaws when it comes to graphical data integrity. Age labels are overlapped and arbitrarily assigned without any explanation. 25 to 34 years of age is used twice for 25 to 49 years as well, and the interval of each age group also varies. They only focused on differentiating colors, not precise and robust labeling. That is, this figure stressed design variation, not data variation. With respect to Annotation and stand-alone readability, this figure has a shortage. There is only a small annotation of the number of countries while there is no clue about standards of income. It is only separated into 4 groups: All, low/lower-middle, Upper-middle, and High-income countries. There is no explanation of income criteria and which countries took part in this telephone interview they mentioned. It makes this figure not able to be understandable itself.

Part 2. Graphic Design

Figure 2

The CO2 Uptake and Concentration in Grass Depending on Treatment and Region



First of all, the main objective of this visualization is to see if there is any distributional difference in CO2 uptake and concentration in the grass, between the 2 regions (Mississippi and Quebec) and the 2 treatments. It is assumed that if the region and treatment are different, the dispersion of CO2 would be different. Also, it is intended to draw a line to indicate a linear relationship within each type of region.

To see the difference in distribution, regarding many 84 plants in the given data set, I chose a scatter plot. It is basically put more weight on positions that are easily decoded. Distinguishable, different 2 colors have been chosen to see the difference between region types. When it comes to operation, the light color of grid lines and black and white theme are used expected to be an anchor to read. The interval of the x-axis and y-axis has been chosen depending on the variation of each variable. Therefore, there is no inconsistency and break in scales.

I also considered the Data-ink ratio. It is likely that only core data is included. There is no chart junk other than the index, annotations, and data point itself. More than 2 dimensions are not necessary, try to equalize the graphical dimensions depicted and data dimensions considering graphical integrity. Less design variation and stress data variation were applied and not given any irrelevant quote.

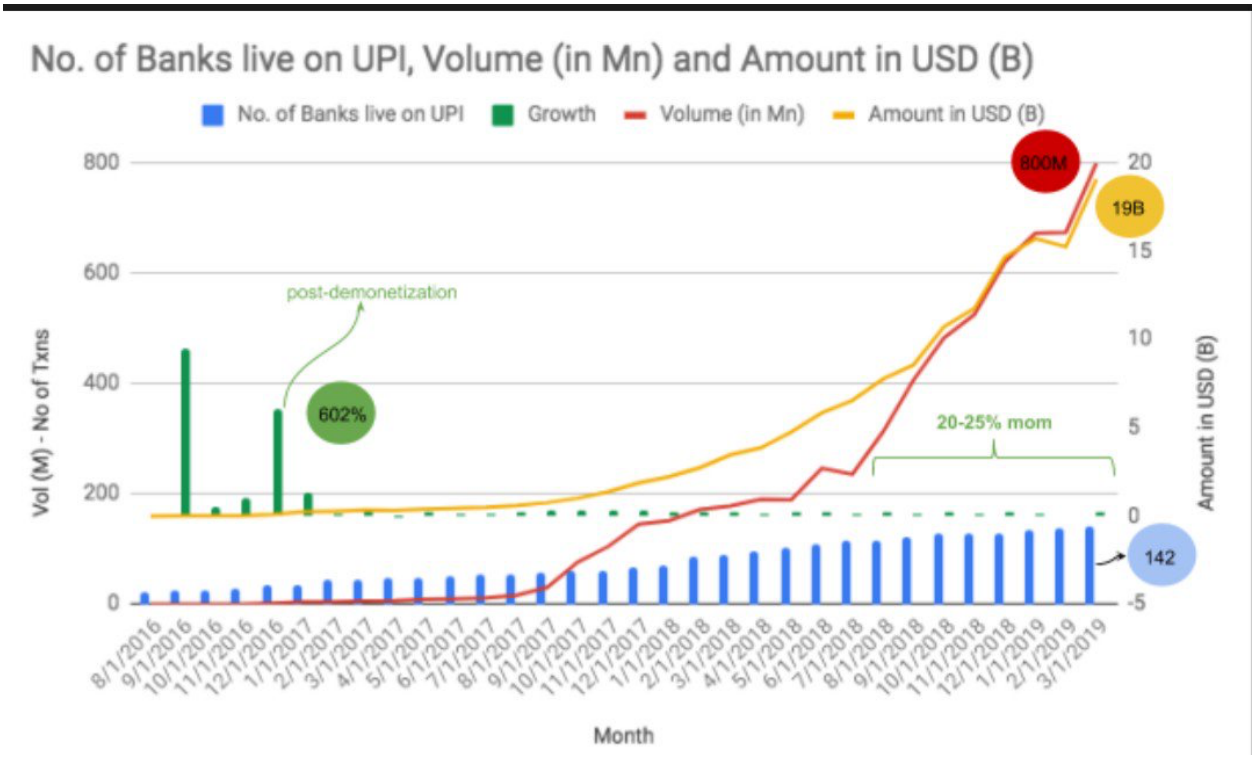
Lastly, Stand-alone readability is not able to be overlooked. To achieve it, I clearly wrote the title and subtitle related to the data. Moreover, the type of regions is separately written on the right side. To inform what those data points are indicating, the x-axis and y-axis are named with units.

Component II: Data Visualization

Part 1. Graphic inquisition

Figure 1

Target Figure to Critique



Note. 2019, Retrieved from https://twitter.com/MohapatraHemant/status/1120212609045655552

1. **Gestalt principles & visual structure**

There are mixed bars and lines in the graphic making readers confusing about what they should focus on. Based on similarity principle, we would perceive the set of green-colored "bars" as a group, while it could also represent the 0-scale level measured by the right vertical axis (dotted line).

2. **Keep it simple: Decoding & Operations**

We can see the trend of lines in the graph, while the bars are not very sensible from both the trend and the values, especially we cannot gain any information about the growth. The grid line is not helpful in this case, as the number of scales in two vertical axes is not equal, and there is even negative scale on the right vertical axis.

3. **Less is more: Chartjunk & data-ink ratio**

The graphic design is too ambitious trying to display values with different magnitudes in one x-y coordinate axis. The bars are redundant in the graphic and should be removed. Instead, simple lines with clear scales and labels would tell the whole story.

4. **Graphical data integrity & lie factor**

As shown in the figure, the highest growth rate is depicted when both the number of transactions and amount of money were extremely low. The intersection point between two lines is also deceiving as we may think the two variables are comparable.

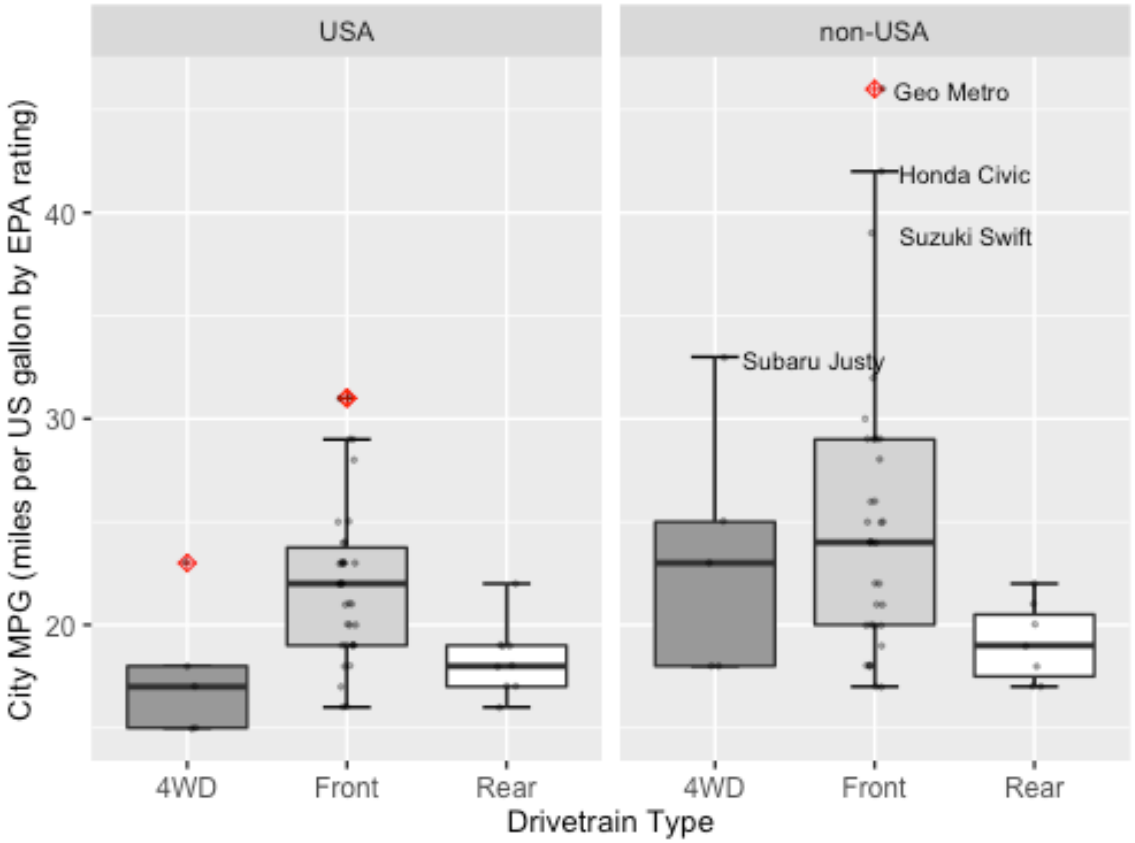
5. **Annotation & stand-alone readability**

The graphic is designed with title, legend, and axis labels. However, with the horizontal axis labeled month, the actual scale is disordered in the format of date, and it is basically not readable. Other marks like "20-25% mom" do not make too much sense. Too much abbreviation involved in the annotation increases difficulties to read.

Part 2. Graphic design

Figure 2

Comparison of MPG of Different Drivetrain Types for USA and Non-USA Cars



Note. The figure is represented by using the *Cars93* dataset in the MASS package in R. In the dataset, 93 cars were selected randomly from passenger car models in 1993 sold in the USA. The two panels demonstrate cars with different drivetrain types which originally from America or not. Four outliers stand out in the category of non-USA cars by using the level of 1.5 times IQR (interquartile range).

Intention and Motivation

The graphic intends to show which drivetrain type of cars is more fuel-efficient among 93 car models made in/outside America. There are three variables represented in the figure, including a continuous variable MPG.city and two categorical variables, drive train type (front wheel, rear wheel, and 4WD) and origins (USA and non-USA). As the graphic illustrated, cars with front wheel drivetrain are more fuel-efficient than those with the other two types. In addition, non-American cars can be driven slightly further per gallon of gasoline compared to American cars for all three types of drivetrain. Four outliers are found in the right panel by extending the greater values of MPG.city to a max of 1.5 times the interquartile range, which help us identify the most fuel-efficient cars in the sample.

The reason why I choose boxplots to represent the data is that we can compare MPGs corresponding to each type of drivetrain independently. The other advantage of using boxplots is to notify outliers (the most fuel-efficient ones or the least ones) easily. Specifically, the same color is applied for each drivetrain type in two facets helps us to sense the differences intuitively. The grid line in the background together with consistent scales assists us read values of five numbers plotted in each boxplot. Jitter is added to show the distribution of observations in each type of drivetrain. In addition, the graph is designed with annotation, including one descriptive title and labels for two axes.

Part 1. Graphic inquisition

Figure 1

The annual incidence of diabetes (age-adjusted)

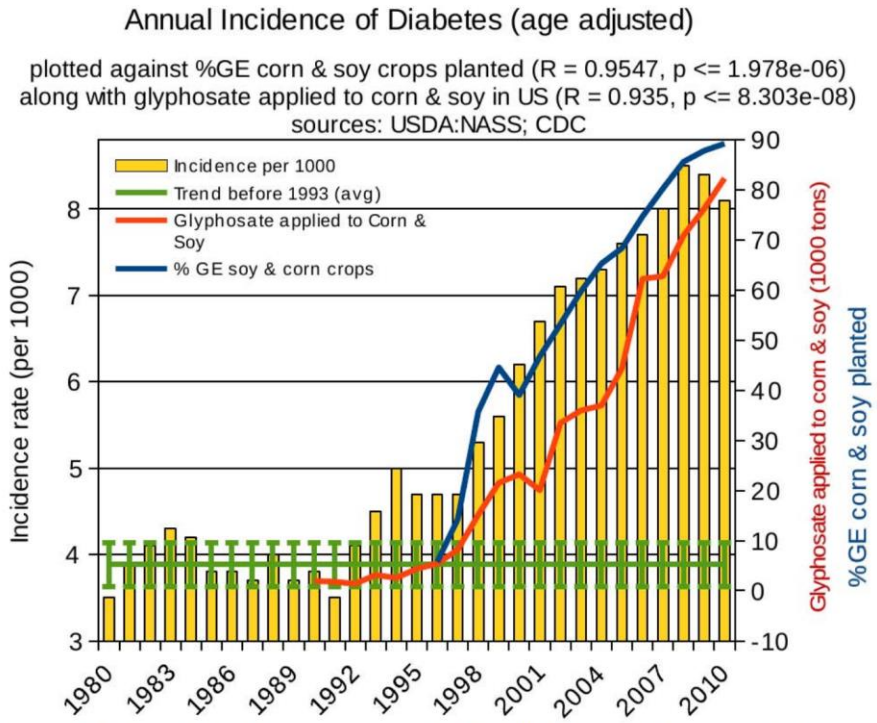


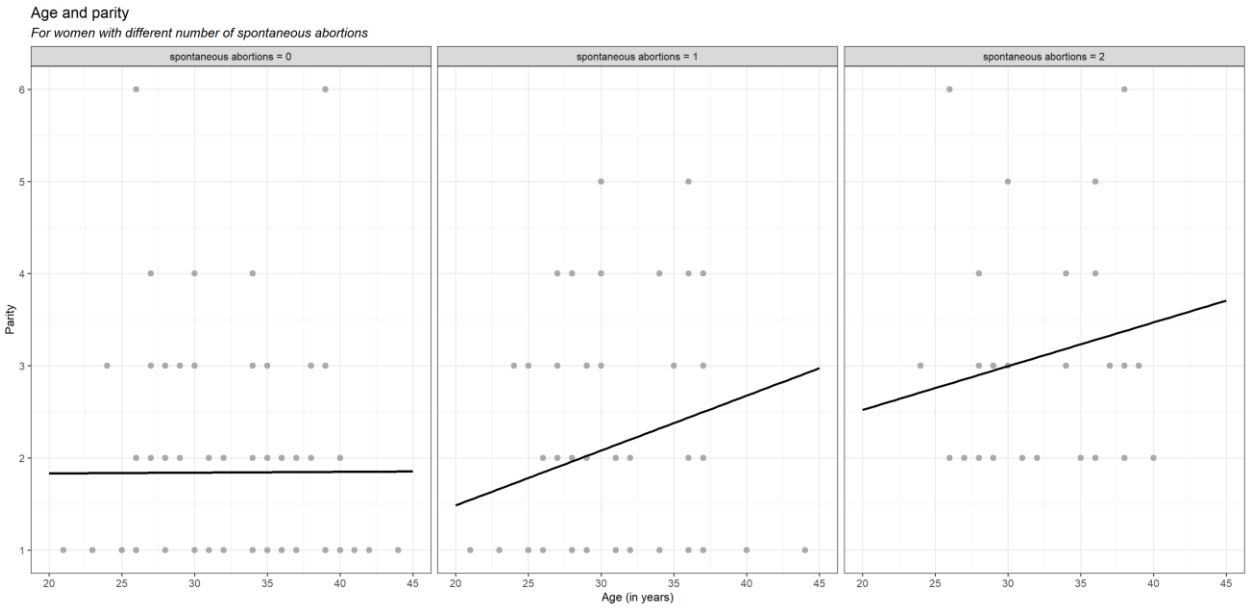
Figure 14. Correlation between age-adjusted diabetes incidence and glyphosate applications and percentage of US corn and soy crops that are GE.

Note: Extracted from Swanson, N. L., Leu, A., Abrahamson, J., & Wallet, B. (2014). Genetically engineered crops, glyphosate, and the deterioration of health in the United States of America. Journal of Organic Systems, 9(2). https://www.farmlandbirds.net/sites/default/files/JOS_Volume-9_Number-2_Nov_2014-Swanson-et-al.pdf

Gestalt Simplicity: There is too much visual stimulation (given by color, shapes and # of elements). Which makes the shapes not simple. Figure-ground: the larger (ground) element is the yellow bars, and maybe the blue and red lines can be understood as figures as they are smaller. For the green lines, is hard to decide if it's figure or ground, probably ground (?) because is as wide as the X-axis. There is also a lack of closure of the blue and red lines. **Decoding** Graph is hard to decode because it contains too much and poorly organized information. Last value in Y-axis in the left is missing (it can be guessed that is 9, but not certain). There are insufficient grid lines to follow the values of the bars, blue line, and read line. Hence, it's hard to make comparisons between the elements. **Data-ink ratio** Number of colors is more than the essential to convey the message. Chartjunk: the small labels inside the graph –which communicates the meaning of the colors- could be outside because it takes away the attention that should be directed to the elements. The labels of the Y-axis on the right repeat what the small label says. Green lines are so thick that obstruct some bar's peaks. Grid lines could be softer to draw attention to the elements. **Integrity** Y-axis on the left doesn't start at 0, which tweaks the perception of the sizes of the bars and lines, making differences look bigger. The Y-axis on the right has been modified to a not feasible -10 value (a % of application - % of glyphosate and, % of GE cannot possibly be 0). This Y-axis modification places the blue line towards the peaks of the bars, somehow visually implying that correlation=causation. **Annotation & stand-alone readability** More annotations than essential are present. However, the green lines seem to express confidence intervals (?) but is not clear, perhaps it could have been represented with a 'position on a scale' line. Is not stated if R is the correlation coefficient and which type (pearson, spearman..). These things diminishes readability. Perhaps, small multiples with less chartjunk could help to clarify the message and comparisons targeted by this graphs.

Part 2. Graphic design

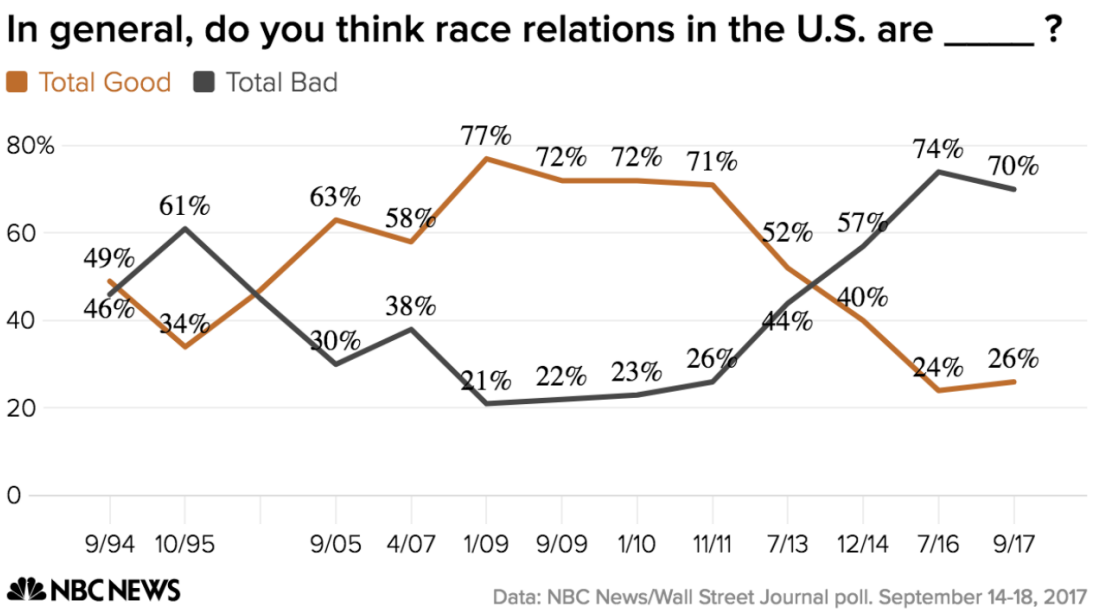
Figure 2



Note: Parity refers to the number of times a woman has given birth to a baby of viable gestation. Disclaimer: This graph intends to be an initial visual exploration of possible patterns. Any claim about relationship between the variables should be supported with solid theory and proper study design.

Argumentation for the design Since my purpose was to visualize possible patterns of two variables across groups (third variable) of women with different numbers of spontaneous abortions, I first created a scatterplot for each group (to observe the location of the research units) and then I added a black line that summarizes the possible pattern in each group. **Gestalt** Similarity: color, shape, and size make the dots look alike, so they can be perceived as part of a group for each facet. Also, the color, shape, and size of the black lines help to group and compare them. Figure-ground: dots are filling a bigger area (than the lines), so they can be perceived as 'ground'. The smaller element across the facets are the lines (compare to the group of dots), so they can be perceived as 'figure'. **Decoding** The selected elements (dots and straight lines) are easy enough for decoding and effective for people to understand the differences. Facets visually help to appreciate that we are dealing with three different groups. Grids serve as references for data points and are equally aligned in all facets. Both scales are consistent (axis X is in line with possible values). The three facets are placed horizontally to facilitate comparison. **Data-ink ratio** The 2 colors used are strictly the necessary amount to represent the graph's message. Grids are soft color and background panel is white to direct the attention towards the dots and black lines. **Integrity** There are two dimensions in both the graph and the data (which is also easier to read). The identical scales and elements used (dots and lines) in all facets + labeling scales, prevents distorting the differences between the facets. The note highlights that visual exploration should not be understood as a 'relationship'. **Stand-alone readability.** The annotations and complete labeling of the 3 variables help to unveil possible patterns between age and parity, and compare them across women with different numbers of spontaneous abortions. Overall, the target pattern is standing out by its visualization through a dark line. Is possible to understand the intended comparison by looking at the graph.

Critique of graph



Source: https://www.nbcnews.com/politics/first-read/nbc-wsj-poll-americans-pessimistic-race-relations-n803446

1. Gestalt principle & visual structure: A lineplot is used to present development over time. In this case angle/direction is a pretty okay way of presenting the data, as it is easy for the eye to tell the change in of the two variables apart.

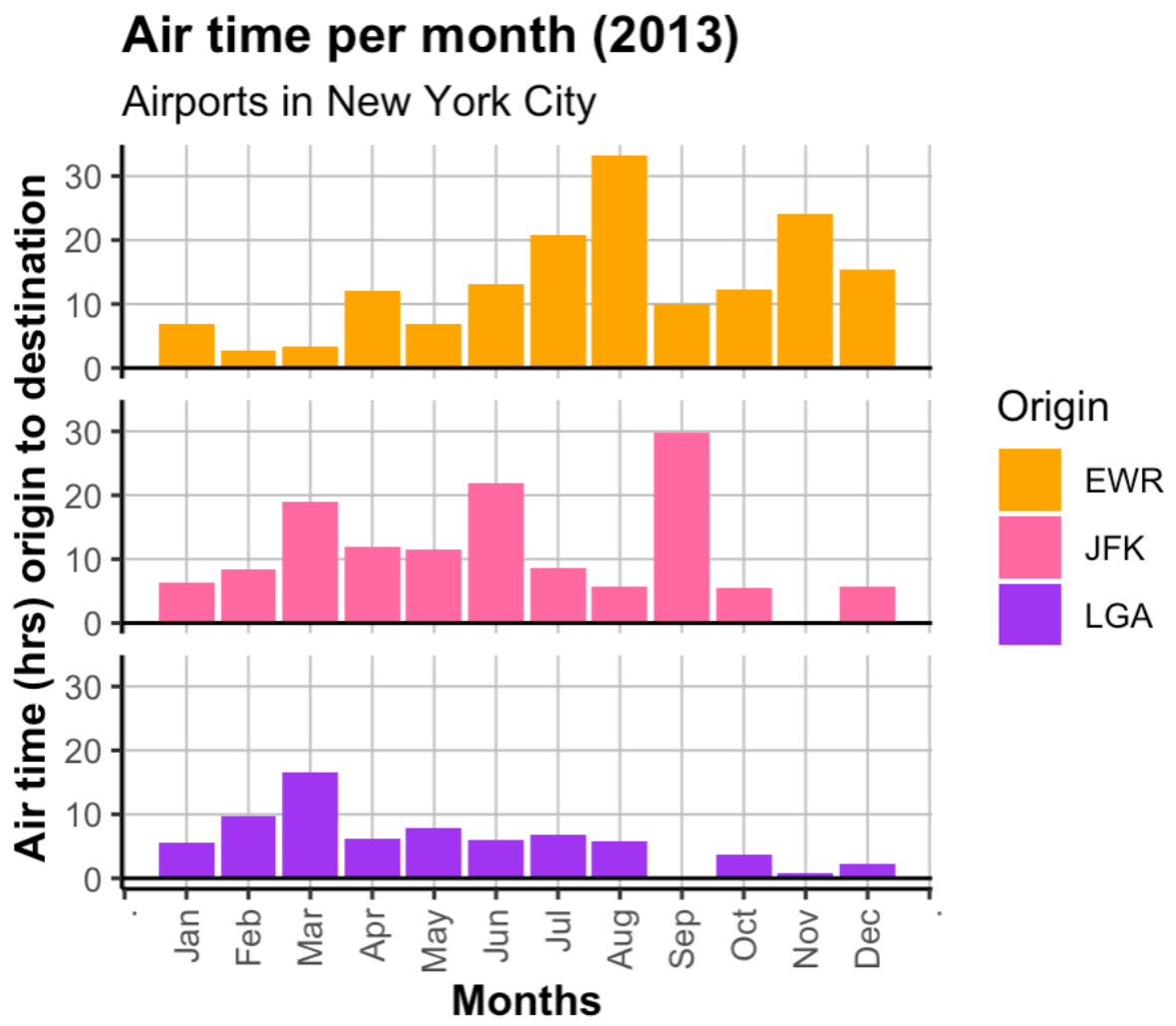
2. Less is more: Chart junk & data ink ratio: The percentages given in the graph make it easy to decode. However I still consider it to be redundant information, as it takes the job of the y- axis. Also, given that the total is 100% for each point of the x-axis, the variables of “total good” and “total bad” will always be opposite of each other (subtracted from 100%), making one of them redundant information.

3. Keep it simple: Decoding & operations: Decoding is made easy by the percentage given for every break on the x-axis. Reading of values without this redundant information would however be hard because of missing vertical grid lines to provide reference and too long intervals on the y-axis.

4. Annotation & stand-alone readability: Stand alone readability is at first glance okay, as the title lets us know a little bit, and one understand that it is some kind of measurement of peoples opinion and a trend showing. However, once you start actuoally looking at the values, it is pretty hard to understand. The annotations are missing for both the x- and y-axis, but we understand that the y-axis shows percentage. Still we are not given any information on what the percentage is of, like who their sample is or the size of it. Is it a random one drawn from the pool of all Americans? Is it only blacks? Whites? Students? Elderly? Voters of a certain political party? We are not provided with much context. In addition, the values on the x-axis are hard to understand. It looks like month/day, but the two first valuesand the text below it looks like month/year. The breaks on the scale are also very strange and inconsistent.

5. Graphic data integrity & lie factor: The sum of each pointon the x-axis vary from 93%-98%, leaving us without information on the missing percentages. The graph follows opinnions over time, so who is in the sample is important to the integrity of the graph. Are they the same or are they different? Is the sample random or biased? Both of these points reduce the graphs integrity.

Self made graph



The figure shows air time per month for three airports. Air time meaning how many hours a plain that left one of the airports spent in the sky before landing at a new destination.

1- Gestalt principles & visual structure: I found the bars to be the best method to show the data in an intuitive and understandable way. I made the bars in color blind friendly collors and put the panels underneath each other to easily compare each month for the different origins.

2- Keep it simple: Decoding & Operations: I changed the scale on the x axis so it hits right on the month and did not cut right in the middle of two months, making it more easy and intuitive to read. I changed the names on labels from numbers to month so it is easy to decode and the reader doesnt have to convert numbers into months. I made the months 3 letters, as it was a lot of text to have whole names. I turned the angle of the labels on the x-axis to get room for all month names so they weren't crammed up and unreadable. I converted air time from min to hrs because hours are easier to understand than tens of thousands of minutes.

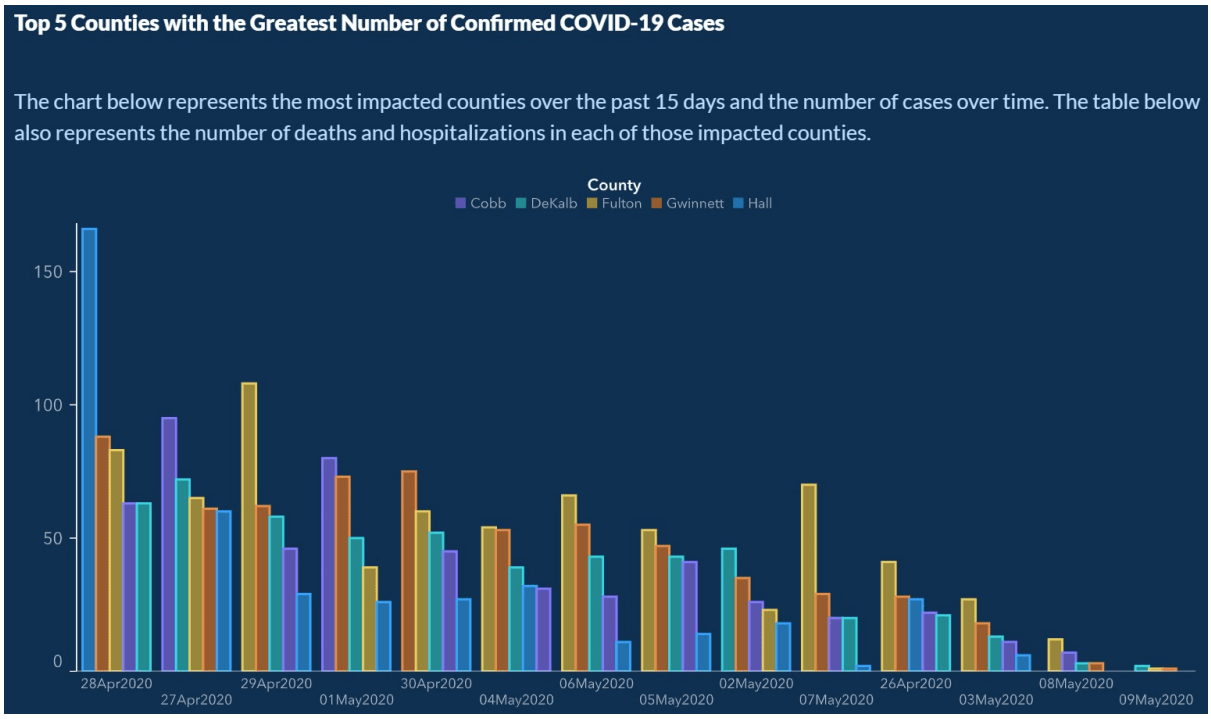
3- Less is more: Chartjunk & data-ink ratio: I made the background black and white so the orofinaly grey one wouldnt be unnecessarily dominating. I also switched out the original grid lines so they were not double but sungular, taking away unnecessary "noice". I removed the facet lables to not have redundant information.

4- Graphical data integrity & lie factor: Tried to use scales that gave integrity, like convert from minutes to hours to give an accurate image of the air time, as 30 000 minutes seemes way more that 30 hours.

5- Annotation & stand-alone readability: I made explaining text and titles bold to make it as explanatory as possible so that the reader would hopefully not be left to any wandering about what the graph is about and having to think har and spend a long time to try to figure it out.

Figure 1

Figure for critique



Note:

The graphic was originally from department of public health Georgia.

https://statmodeling.stat.columbia.edu/wp-content/uploads/2020/05/image.png

Gestalt principles and visual structure. There is a clear visual separation in the graph. It uses Y and X-axes to bring about enclosure and define the data space. Similar elements are grouped together, i.e., each bar represents a county. County cases are sorted and grouped together so that within each cluster the cases rank from highest to lowest. This creates an impression of a downward trend but is more of an illusion since the days are not ordered chronologically.

Keep it simple: Decoding and Operations. The chart is simple to perceive, but the choice of graph is wrong. Graphs with clustered columns are not the best for depicting a trend which I believe is the intend here, a better choice is a line plot using small multiples either by means of faceting or panels. The strange ordering of days e.g., 28th April following 27th on X-axis and the resorting of the counties each day makes it hard for the eye to follow the trend.

Less is more: Chart junk and data-ink ratio. Drop the blue background as it codes for no additional information and is unrelated to the data. Similarly drop the bolding thought out e.g., in the title. Drop the two text lines below the title, these are a bit redundant. It is also better to sort the chart by data attributes rather than county names.

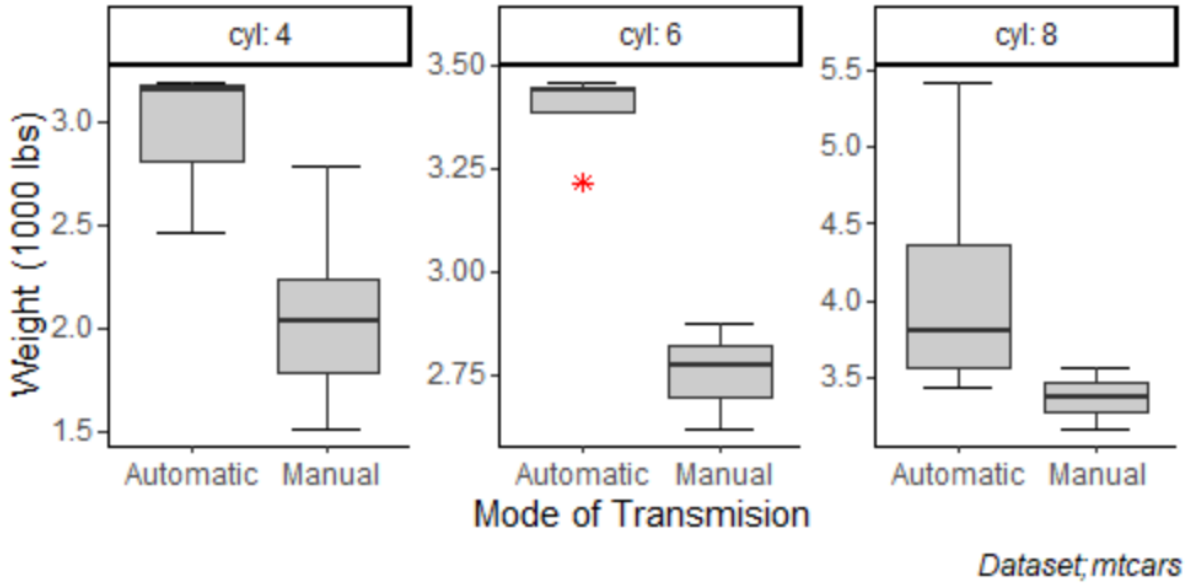
Graphical data integrity and lie factor

Truthful graphics have a lie factor of 1 i.e., the size of effect shown on the graph is directly proportional to the size of the effect in the data. A look at the data set shows that new cases for Cobb and Dekalb are 56 and 52 respectively but on the graph, they are shown as equal creating a distortion. Additionally, the impression of downward trend over time is an illusion as measures themselves are not ordered chronologically across time on the X-axis.

Annotation and stand-alone readability. The Y-axis lacks axis title and units of measurement which makes it hard to know what the numbers represent. X-axis lacks also lacks axis title. An informative title and legend is present which compensates a bit for this lack but only implicitly and not good enough for standalone readability.

Figure 2

Distribution of Car Weight by Number of Cylinders and Mode of Transmission



Note:

Through box whisker plot I show the spread of the data by factor. Generally, the chart shows that automatic vehicles have higher weight compared to manual vehicles. Similarly, the graphic suggests that the higher the number of cylinders a vehicle has, the higher the vehicle weight so that vehicles with four number of cylinders weigh the least in the dataset. Cyl: means the number of cylinders a vehicle has. Note also that six-cylinder vehicles has some outliers. The data was extracted from the 1974 Motor Trend US magazine and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models) and is freely available in R program.

https://www.R-project.org/

Box whisker plots provide a lot of information in a single chart i.e., median, outliers, minimum, maximum values, lower and the upper quartiles, distribution, and skewness. I am interested in exploring the associations between various aspects of car design on the weight, specifically the effect of the mode of automobile transmission and the number of cylinders on the weight of the car as measured by weight per 1000 pounds. Additionally, the weight variable in my dataset is a continuous variable whereas the number of cylinders and the mode of transmission are categorical variables. All these coupled together makes boxplots a better choice for visual exploratory statistics.

Gestalt principles and visual structure. There is a clear visual separation in the graph. I use Y and X-axes to bring about enclosure and define the data space. Similar elements are grouped together, i.e., each boxplot represents a category.

Keep it simple: Decoding and Operations. The chart is simple to perceive. I sorted the chart by data attributes rather than car names. Dataset has 11 variables & 32 observations, but I have used 3 variables to tell my story. The small multiples help visualization.

Less is more: Chart junk and data-ink ratio. I dropped the background and the grid as they coded for no additional data. Colouring each category is unnecessary as the chart labels are enough to show their differences, so I kept the boxplots colour uniform. Likewise, no need for a separate legend as the labels are in the chart.

Graphical data integrity and lie factor. Truthful graphics have a lie factor of 1 i.e., the size of effect shown on the graph is directly proportional to the size of the effect in the data. In the dataset mean weight for 4-cylinder automatic cars is 2.04 whereas in the box plot the median weight for 4-cylinder automatic cars is about 2 which is almost directly proportional.

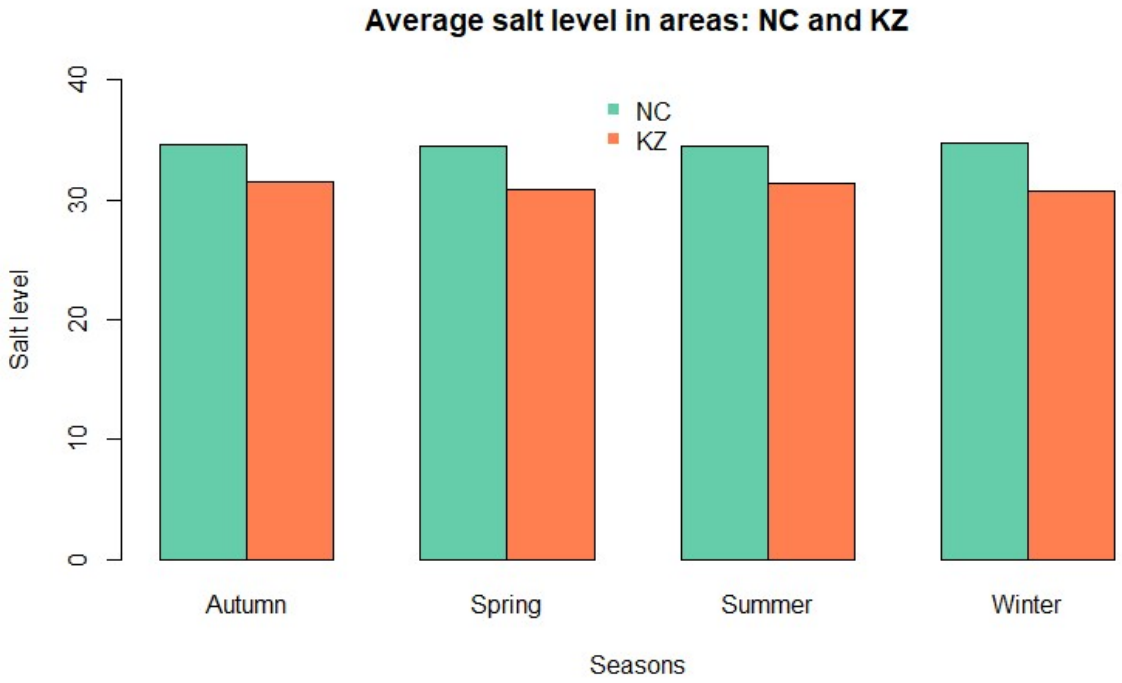
Annotation and stand-alone readability. Y-axis title and well labelled Y-axis with units of measurement brings meaning to numbers. X-axis has labels and title. Labels are within the chart to help standalone readability and avoid the need for a separate legend.

Part 1. Graphic inquisition

Source: https://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6?r=US&IR=T

1. Gestalt principles & illusions. Logos of the fast-food restaurants is not clear, especially the ones that are too small. and nothing is written in the U axis. However, the difference of right and left side is clear.
2. Keep it simple. Decoding & Operations. The figure is relatively simple to perceive.
3. Less is more: Chartjunk & data-ink ratio. However, the map and GDP of Afghanistan is creating more confusion than giving some information to compare the sales values.
4. Graphical data integrity & lie factor. The size of the logos and the sales amount is not accurate. For example, the sales number of McDonalds is about 4 times compared to that of Burger King; but the ratio of their logos' area is about 8 times. It seems to imply a trend of sales across various fast-food restaurants, but the size of the logos misleads about actual differences in sales.
5. Annotation & stand-alone Readability. The values in Y-axis is given, but has no label. The label is repeated several times ("billion in sales) inside the graphic. The label of X- axis is missing. Lacks title of the graph. There is no reference of time.

Part 2. Graphic design



Variables used to generate my own figure are: Mean salt level, Area, and Seasons.

1- Gestalt principles & visual structure. Clear bar chart showing Salt levels across the 4 seasons of the year in two areas, NC and KZ.

2- 2- Keep it simple: Decoding & Operations. Simple 2D graphs have been used. 3 variables were not shown in a clumsy 3D diagram.

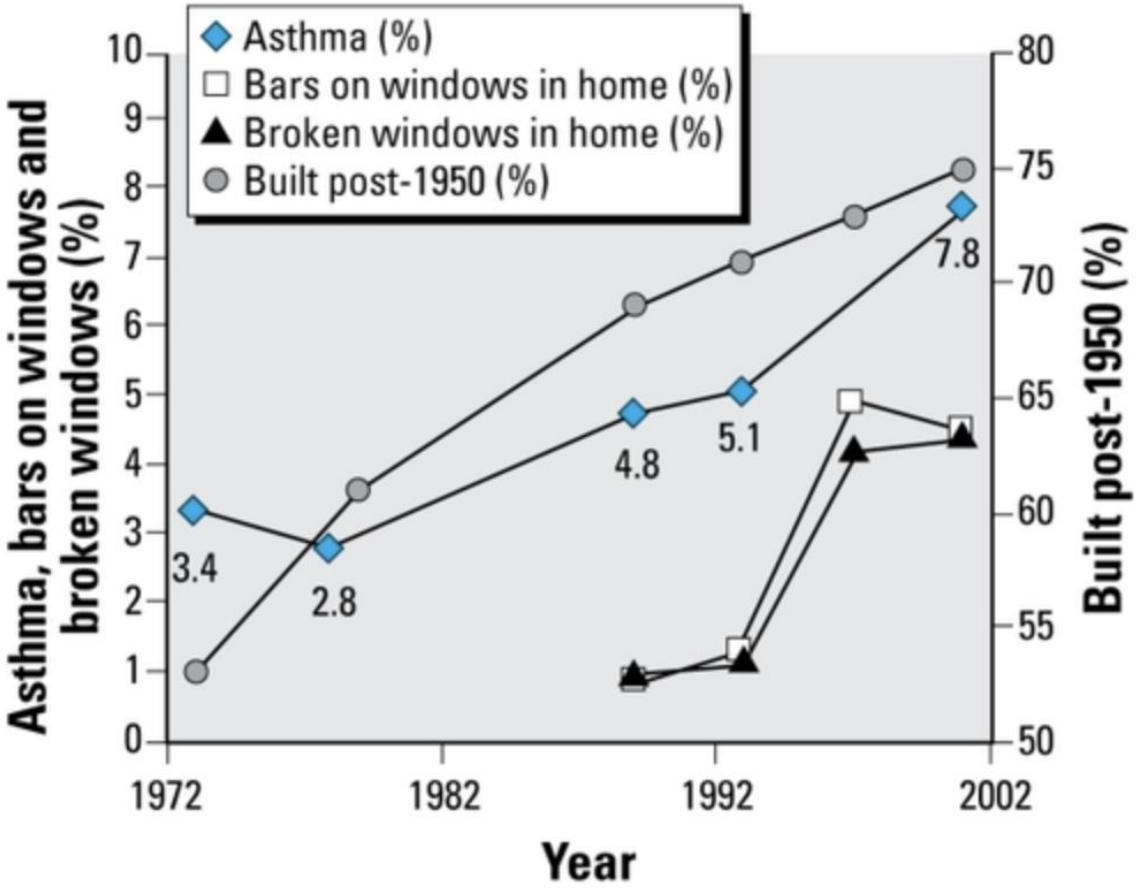
3- Less is more: Chartjunk & data-ink ratio. Light, but contrast color is used for the two areas.

4- Graphical data integrity & lie factor. Data did not have information on which year's data this. It would have been useful to put this information the graphic.

5- Annotation & stand-alone readability. The axes are clearly labeled. It clearly shows that on average in every season NC has higher salt levels than KZ. In both areas, there are seasonal variations as seen by the different height of the bars.

Figure 1

Changes in asthma, bars on windows, broken windows, and year built over time.



Note. From "The Relationship of Housing and Population Health: A 30-Year Retrospective Analysis." (p. 600) by D. E. Jacobs et. al. 2009, Environmental Health Perspectives.

Source for graph

Jacobs, David. E., Wilson, J., Dixson, S. L., Smith, J., & Evens, A. (2009). The Relationship of Housing and Population Health: A 30-Year Retrospective Analysis. *Environmental Health Perspectives 117*(4), 597-604. https://doi.org/10.1289/ehp.0800086

Gestalt principles

The use of lines between the data point in the Figure 1 shows the continuity. The datapoint that are from the same categories are also presented in the same colors showing similarity. The different variables are also separated with the use of shapes.

Decoding

The use of two axes representing percentages makes Figure 1 more challenging to decode. Both are on a scale of percentages, but the range of the percentages are different. Using only one range of percentages would make Figure 1 easier to interpret.

Less is more

Figure 1 only includes data point or elements that makes the graph more easily understandable. The only possible exception is that it is not necessary to include both differing shapes and colors to separate the variables, it would probably enough to just use colors. There is no reason for including numbers on the asthma values, more than adding this to the others.

Graphical integrity and lie-factor

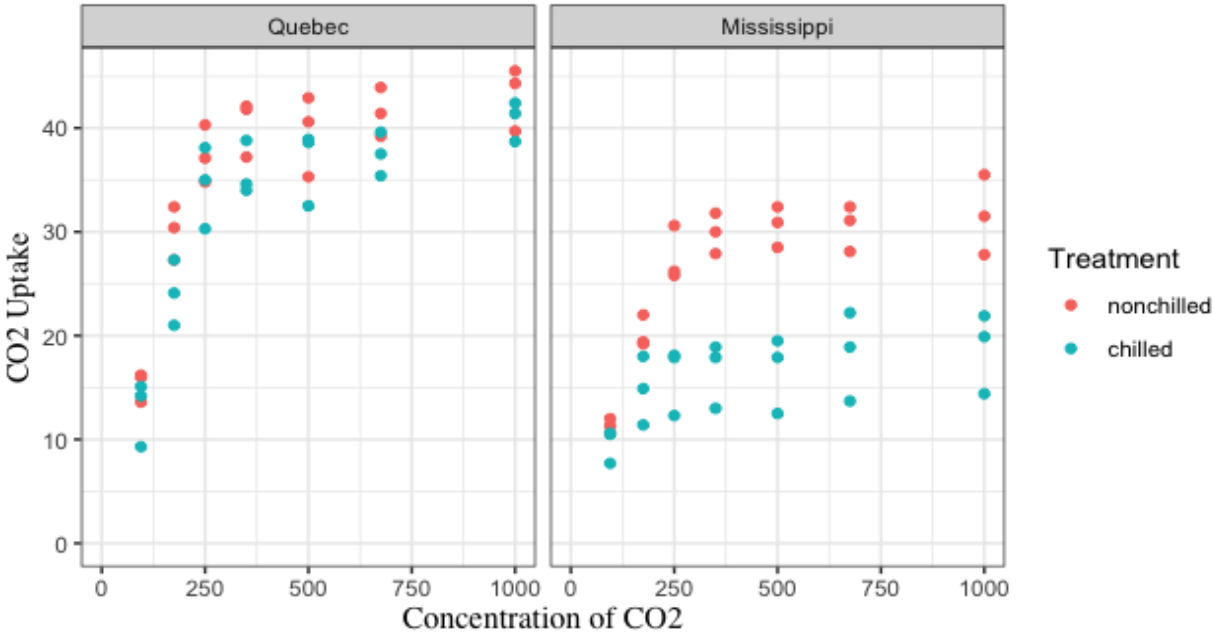
Figure 1 includes two y-axes, which represent different ranges of percentages. The lefts range is 0 to 10 percent, while the right has a range from 50 percent to 80 percent. This makes it look as though the changes over time for the four variables are of the same magnitude. This is misrepresentation of the actual changes. If the variables were presented on a graph with a zero-point, it would become clear that the variables differ more than what the graph presents.

Annotations and Stand-alone readability

The y-axis title on the left is helpful in understanding which datapoints belong to the left scale and which belong to the right. The label on the right y-axis does not make the variable belonging to that axis clearer. The title of the graph does not help explain the motivation of the graph of what it is meant to make clearer. A note should have been included to clarify the graph.

Figure 2

Scatterplot of CO2 Uptake of Plants on Different Levels of CO2 Concentration, by Origin of the Plants and Type of Treatment.



Note. This figure shows how much Carbon Dioxide the plant *Echinochloa crus-galli* takes up on different levels of concentration of CO2 in the air. The points are colored according to the received treatment.

Figure 2 intends to show how much Carbon dioxide the grass-plant *Echinochloa crus-galli* takes up at different levels of concentration of CO2 in the air. It is shown with two variables separating the plants. These variables are their origin, either Quebec and Mississippi, and types of treatments administered, chilled and nonchilled. The scatterplot is chosen as it makes it possible to show how the CO2 uptake develops on the different CO2 concentration levels. The colors were chosen to illustrate and make clear the difference between the plants that were chilled and the plants that were not chilled. The colors red and blue are chosen on purpose as they give associations to the cold and warm. Blue gives association to cold, while red gives association to warm, similarly to the chilled and nonchilled treatments of the graph. This can both make understanding the graph and differentiating between the two values of the treatment variable easier.

The two facets illustrate the two possible origins of the plants. They are shown separate in two facets as to illustrate that there is a clear difference in the CO2 uptake of the plants from the two regions.

Literature

R: Carbon Dioxide Uptake in Grass Plants. (n. d.). 13. oktober 2022, from https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/zCO2.html.