Data Visualization

Anni Elisabeth Enevoldsen
aenev20@student.sdu.dk

Benjamin Mikael Andersen
beand19@student.sdu.dk

Andreas Edal Pedersen
anepe19@student.sdu.dk

Svend Anton Vibæk
svvib19@student.sdu.dk

University name - University of Southern Denmark, SDU

Institute - Mærsk Mc-Kinney Møller Instituttet

Course code - T520040101

Hand-in date - Sunday, December 11th, 2022

Group nr. - 4

Dashboard link - https://benji703.shinyapps.io/BestSellingAlbumsByDuration/

Abstract

The dataset “Best Selling Albums By Duration (1990-2021)” chosen for this project is originally sourced from bestsellingalbums.org. With this data, the group intends to create graphs that will allow the reader to get more familiar with the most popular albums, artists, and genres throughout the years.

The questions formulated were “*Q1: How does the development of album sales and the year correlate?*”, *“Q2: Which genres sell best throughout the years?”*, *“Q3: How many times are artists repeated?”*, *“Q4: How are the sales distributed between the genres?”*, and *“Q5: What artist is recommended if the reader prefers a specific genre?”*.

To best find and convey the results, the group has made different graphs that answer the questions. The results gained from making the graphs show that:

Album sales have been decreasing throughout the years *Q1,* and in the beginning rock was the best-selling genre, however since 2008 it has been pop *Q2*. Most artists on the list have only had 1 album, however the artist with the most albums on the list has made 14 albums *Q3*. There is not a big difference in how the sales are distributed across genres when the amount of entries are taken into account *Q4*, and the highest-selling artists in the genre chosen by the reader will be recommended *Q5*.

These results are shown in graphs that can be seen in the dashboard.

Table Of Contents

Background & Motivation

The music industry has changed a lot during the last 30 years [1], which is why it would be interesting to see how the top albums have changed.

To explore this, a dataset consisting of the 10 best-selling albums per year for the last 30 years was chosen. The dataset contains 12 attributes, each providing different information about the albums, spanning: the year of release, album name, artist, and album length, along with other values. Each attribute can help create visualizations of the trends and development of album sales, in order to assist the viewer in getting an understanding of these aspects.

Project Objectives

The questions formulated for visualizing the dataset to improve a reader's understanding are shown below.

- **Q1: How does the development of album sales and the year correlate?**
- **Q2: Which genres sell best throughout the years? (and total)**
- **Q3: How many times are artists repeated?**
- **Q4: How are the sales distributed between the genres?**
- **Q5: What artist is recommended if the reader prefers a specific genre?**

With these questions answered a lot of knowledge about the album trends throughout the last 30 years can be gained.

Data

One dataset that can answer the above questions is the data set describing the best-selling albums from 1990 to 2021 [2]. It consists of the top ten albums that year with additional information.

Year	Ranking	Artist	Album
Worldwide Sales (Est.)	CDs	Tracks	Genre
Album Length	Hours	Minutes	Seconds

Variables present in the data set are year, ranking, artist, album, worldwide sales in whole units, number of CDs, number of tracks, genre, and album length, both in hours, minutes and

seconds. As it is the top ten albums per year and it covers 32 years, the total number of items is 320 each representing an album. The variable types of the variables/attributes can be seen below.

Variables	Variable type
Year	Numerical discrete
Ranking	Categorical ordinal
Artist	Regular categorical
Album name	Regular categorical
Worldwide Sales in units	Numerical discrete
Number of CDs	Numerical discrete
Number of tracks	Numerical discrete
Genre	Regular categorical
Album length in hours	Numerical continuous
Album length in minutes	Numerical continuous
Album length in seconds	Numerical continuous

No data cleanup is needed for the data set as it is small, well-organized, and contains no noisy data or null fields.

Visualization/Dashboard

Audience

The data contains basic information about the top hits within music. The main audience of the data is common people who would like to know more about music. The dashboard will therefore be structured so that readability and ease of access are the top priority.

Dashboard Layout

Readability and ease of access will be achieved by using a long page, like a website, structured as a story.

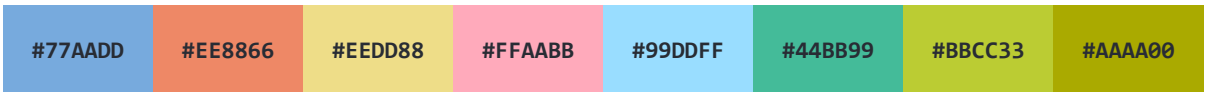
This fits the casual reader that just wants ease of access. There will be no menus or tabs as this might turn off the reader from exploring further due to these necessary actions.

The data will be presented as an *author-driven* and *reader-driven* story. First, the author has control and then the reader can explore an interactive graph at the end of the dashboard.

Design

Color

To take accessibility into account, the first thing to be done is to choose a color theme that provides 8 colors [4] distinguishable by people suffering from the most common kinds of colorblindness. Many color schemes have been created and there is therefore a large variety to choose from. The group has chosen to go with Ph.D. Paul Tol's "light" qualitative color scheme [3] as it is pleasing to the eye and gives off a modern look while still keeping all the essential properties of an accessible palette.



Font

The font choice plays a big role when communicating the information of a graph efficiently. To ensure the best possible viewing experience on a digital display, a sans serif font will be chosen as it is the most readable font type on smaller displays and resolutions. The distance between the letters will not be changed as the default font settings are made to give the best common viewing experience. Bold text will be utilized to highlight important parts of the graphs and separate headings from normal text.

Shapes

The graph will make use of squares, circles, along with other polygons as long as they make sense for the specific graph. When making points, the main choice of shape will be circles as they are easier on the eyes and provide a streamlined and modern look.

Q1: How does the development of album sales and the year correlate?

Attributes: Worldwide sales (numerical: discrete), year (numerical: discrete)

- **Visual channels:** Length, Area, Color
- **Multivariate chart:** Total album sales (quantitative), year (quantitative) and grouped by genre (categorical)

This graph aims to show the change in worldwide sales over time. As the data is quantitative the charts should use magnitude channels like position and length as it is easy to read.

To incorporate an extra dimension the sales can be divided by genre per year using a stacked area chart. This introduces categorical data as the genre. The visual channel should now be an identity channel to tell the reader which category the given data stems from. This channel can be the color of its section in the area chart.

The number of values must be reduced as 10 unique genres exist through the records in the genre attribute. This is too much as the chosen color palette only consists of eight colors. This introduces the need for an “other” category for the genres where the least-selling genres are combined.

Annotations help the reader understand sudden changes in sales and get an overview of other important events. These do not conclude any causation of the change in sales but rather give the reader the means to reflect on the data.

The area is used to display the total sales of a given genre but this is not the primary purpose. The primary attribute to visualize is the total album sales across all genres per year which are displayed on the y-axis as the height/length of the stacked graph. Length is better than area according to Steven’s Psychological Power Law [5], but it is good enough as the main purpose is not the distribution of the genres but the overall development of the total album sales.

A stacked bar chart has also been created to visualize the same data. This better communicates that the important data is the total sales per year (the length) and not the total sales across all years (area). The bar chart can be seen in figure 1.

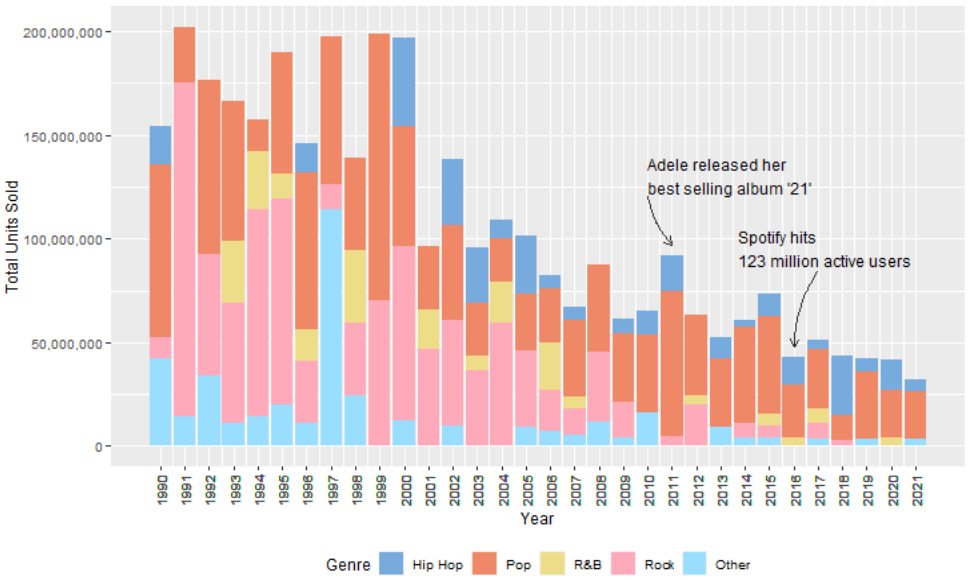


Figure 1: Stacked bar chart showing the total sales per year

Q2: Which genres sell best throughout the years?

Attributes: Genre (regular categorical), worldwide sales (continuous), year (discrete)

- **Visual channels**: Color (hue), Tilt/angle

The goal of this question is to make an animated and interactable graph where the reader can toggle the genres of choice and watch the change in sales throughout the years. To achieve this purpose, gganimate was the initial idea, however, since it did not allow for interaction, the whole graph was made using R-Shiny. The idea of the graph was to have the year on the x-axis and the number of albums sold per genre on the y-axis. Each toggled genre would then be represented by a separate color on the graph. Since the dataset has more genres (10) than there are colors available (8), the genres with the smallest amount of total sales have been combined into a new category called “Other”. This ensures that there is only a maximum of 8 possible colors in the graph.

To animate the graph, a slider is made to represent the newest viewable year. This slider can then be animated by adding the `animate = TRUE` property to the `sliderInput`. Pressing the “play” button will now incrementally increase the year slider by 1 every second.

A list of checkboxes is made to allow the reader to filter the genres of choice so that the graph can look as they want. These checkboxes are initially defaulted to the 4 most popular genres for simplicity.

Since the graph is reactive, when a checkbox is checked, the whole graph is updated. This means that the new genre will appear on the graph immediately and the legend underneath the graph will automatically update to show the new genre and its color appropriately.

All of the interactive elements have been centered to make it an easier viewing experience as it allows the information to be closely connected visually.

With the viewing experience in mind, the y-axis, representing the number of total sales, has been split into Millions. This is primarily done to make it easier for the reader to quickly decipher. Another reason is the fact that the original number is an approximation and the exact number, therefore, isn't as important.

A label is created on the most recent year to display which genre it is and the amount of combined album sales that year for that genre has also been made. The label is filled with the same color as the line on the graph to make it easy to correlate the two.

Lines are used between the year points to guide the viewer's eyes from point to point and thereby ensure a better viewing experience.

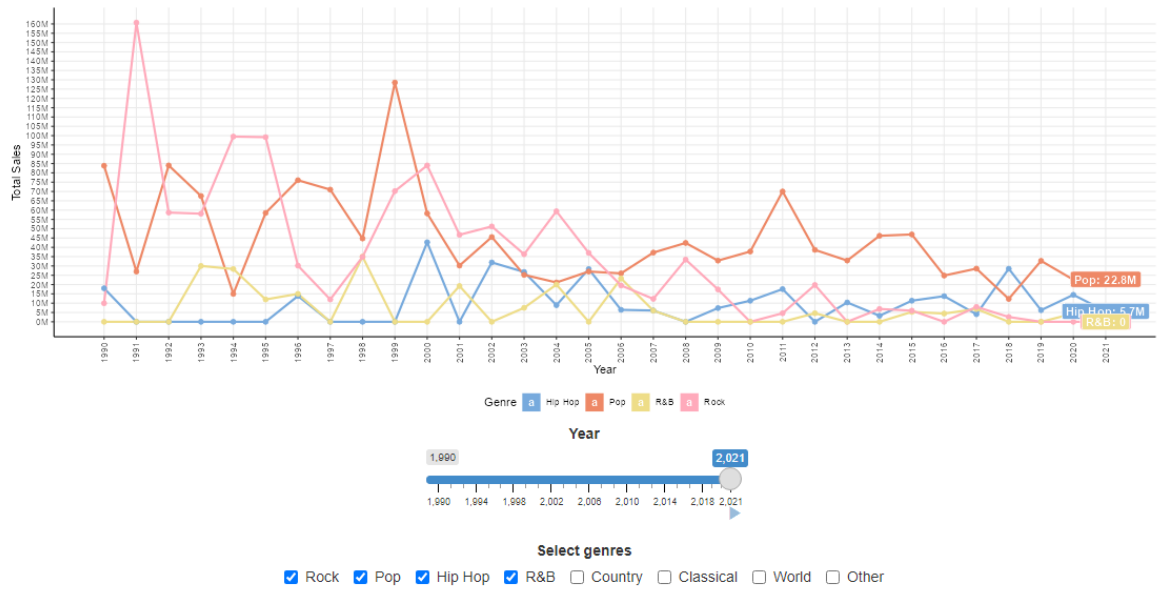


Figure 2: Interactable and animated graph - Which genres sell best throughout the years?

Q3: How many times are artists repeated?

Attributes: Album count(numerical: discrete), artist count(numerical: discrete)

To show this, a bar plot and a tree map will be used. Below, the bar plot can be seen, with the number of albums an artist has released on the x-axis and the number of artists on the y-axis.

The bar plot shows the number of albums each artist has made that have made it to the top 10. Here it is easy to see that most of the artists on the list have only released one album that has made it to the top 10, and that there is at least one artist who has released 14 top 10-selling albums.

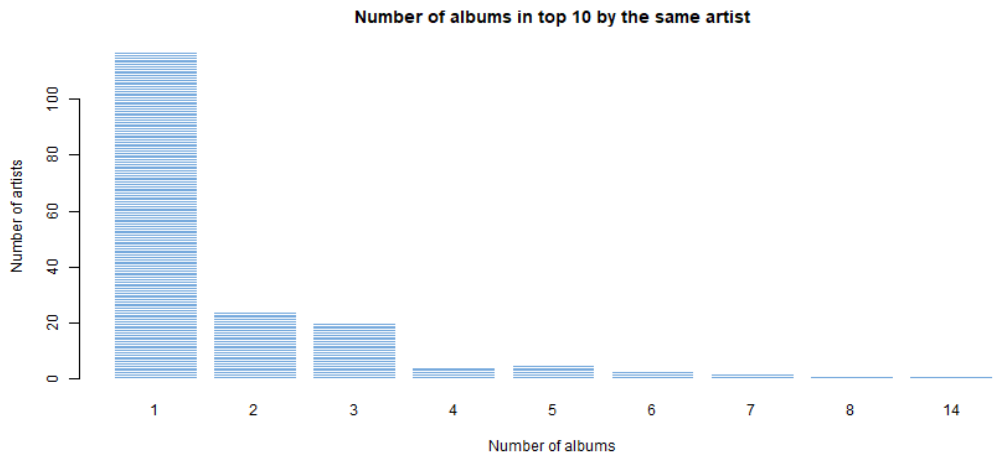


Figure 3: Bar plot - How many times are artists repeated?

Attributes: Attributes: Album count(numerical: discrete), artist count(numerical: discrete)

- **Visual channels**: Color (hue), Size

The treemap is used to show the same data in a slightly different manner. With this graph, it can be easier to see the results and compare them to each other. The treemap can help visualize the difference in proportions better. It is also great for readers who are more visual. Both of the graphs together can help support the reader to gain a better understanding of the data and the results.

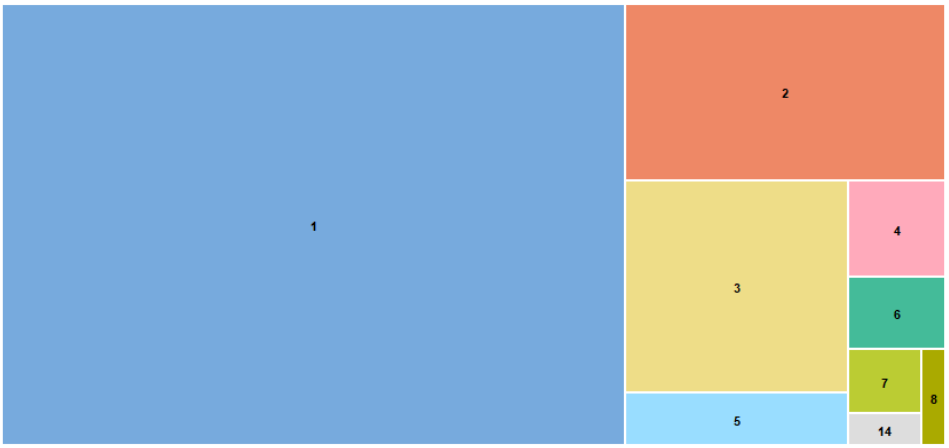


Figure 4: Treemap - How many times are artists repeated?

Q4: How are the sales distributed between the genres?

Attributes: Genre, Sales

- **Visual channels:** Color, length
- **Bivariate chart:** Genre: regular categorical, Sales: numerical discrete

This data will be displayed through a violin chart that shows the density of the sales divided into the different genres. A boxplot will be added to show the quantiles, the mean, and any outliers. This will help to compare the different distributions and will help the reader understand the violin chart and its data.

An “others” category is also created to minimize the number of categories in the genres attribute. This is again to reduce the number of colors used on a categorical attribute with the aim to make it easier to read and understand.

Color is added to separate the different genres even more. This is a redundant visual channel that can cause confusion for the reader but it is used to keep the same theme as in Q1 and Q2.

Q5: What artist is recommended if the reader prefers a specific genre?

Attributes: Album, Artist, SalesByArtist, Sales

- **Visual channels:** Length
- **Bivariate chart:** Artist and albums: Regular categorical, Sales: numerical discrete

To answer this question the dashboard needs some user interactability. This is done by having a dropdown menu, where the reader can select their preferred genre from a list. To give the reader some insight into their selected genre the dashboard will have two horizontal bar charts, showing them the most sold albums and the artist with the most sales respectively. The artist and albums will be represented on the y-axis, while the sales are on the x-axis. As there are 320 albums, the graphs will be limited to showing the top ten entries for each category.

The reason for making these charts interactive is that the dashboard tries to give recommendations to the reader, based on their preferences. Another factor is that, even though the dataset only consists of 320 items, that would still be too much data to show in a single bar chart, hurting the readability of the visualization.

The only visual channel used in these visualizations is length, as this is the most effective visual channel to gauge the difference between data points [5]. This means it will be easier for the reader to see who has the most sales and by how much. Another aspect that increases the ease of reading is the fact that the charts are bivariate, making them more simple to read. The variables of artist and album are regular categorical, while the sales variable is numerical discrete, as it refers to copies sold. This makes it fit well into bar charts, as there is no order to the categorical variables, which is for example needed in a line graph.

Finally, to accompany the bar charts, an interactive table is added. This will show all the albums of the current genre with the option to go through the pages and sort them based on attributes such as: year, length in minutes, and sales, all of them both in ascending and descending order. As these visualizations aim to highlight the best-selling albums, it is sorted by sales in descending order by default.

Must-Have Features

Of course, the project must follow the requirements given in the course, but there are also several features that the project group would like to incorporate into the dashboard to make it more complete. Following is a list of the requirements and the must-have features:

Requirements:

- The dashboard must have at least 8 graphs
- There must be at least 3 different types of graph
- The dashboard needs to have at least one animated graph
- A download link to the report must be present in the dashboard
- A link to the dashboard must be present in the report

Must-have features:

- The dashboard should have at least one interactive graph
- The charts in the dashboard must be colorblind friendly

Optional Features

The following list shows the optional features of the project:

- The reader should be able to apply filters to the data
- The reader should be able to sort the data based on its attributes
- There should be at least one graph with semantic zooming

- Annotations should be used for at least one graph

Story/Results

The general story of the dataset can best be described through the first question: "*Q1: How does the development of album sales and the year correlate?".* From the beginning, the project group had a pretty good idea of the development of album sales, as they followed the move from physical albums to digital albums and then to music streaming. Therefore the expectation was that the total album sales had gone down through the years. It was apparent that the expectations were right when the graphs for *Q1* were created. Here you can see a steady decline from the 1990s, where total album sales of the top ten albums could peak at over 200 million copies, while the peak in the last five years is around just 50 million in 2017. So the answer to *Q1* is that album sales have been decreasing throughout the years.

The results of the next question: "*Q2: Which genres sell best throughout the years?"* came as more of a surprise to the group. The expectation was that pop was consistently at the top with rock being close by most of the time at least in the 1990s. Though, it was seen on the visualization that rock was the dominating genre through multiple years through the 1990s, while even achieving a few victories into the early 2000s. It is prudent to keep in mind that this dataset only consists of the top 10 albums from each year. The effect of this is that single albums can change the results drastically, as a search through the data made it apparent that rock's biggest victory in 1991 was helped by both Metallica and Nirvana releasing an album that year. To give the reader a better understanding of this correlation, it could be annotated on the graph or highlighted in another manner. After 2008 pop has consistently stayed on top of album sales with the sale of rock being diminished to its lowest point ever.

"Q3: How many times are artists repeated?" is interesting, as the results show that about a third of the artists who have made a top 10 album have actually made 2 or more that were top 10, while the most popular artist, Taylor Swift, had 8 albums that made it on the list. Another artist was listed as having made 14 top 10 albums, however, this is not an actual artist, but albums made from different movie soundtracks.

"Q4: How are the sales distributed between the genres?" is answered with a violin chart with boxplots inside the "violins". Before plotting the chart, the project group expected the distributions to be skewed towards the bottom, with a few outliers at the top. This

expectation was based on a look at the dataset that revealed that there were a few albums with much higher sales than the others. The chart ended up being close to what was expected, with most of the genres having a few outliers at the top. The *R&B* genre did, however, not have these outliers, but that could be due to the limited number of *R&B* albums in the dataset. Having more albums might show a similar distribution as in the other genres. As each genre's area is normalized, it might give the reader the wrong idea and make them think that all the genres are equal in sales. This could be mitigated by making the area related to the actual amount of sales, but this could also hinder readability as some of the "*violins"* would be much larger than others.

When answering the question: "*Q5: What artist is recommended if the reader prefers a specific genre?"* one main assumption is made, namely that the artists people will like are the ones who have sold the best. The expectation when formulating this question is correlated to this assumption, in that the graphs will show the best-selling artists. The graphs used to answer this question are two bar charts, showing the top 10 best-selling artists and albums in regard to the selected genre. The main thing learned from these visualizations is that some genres are quite underrepresented, as they only have a few albums or even just a single album to represent it throughout the entire 32 years. With this in mind, it would be interesting to look at a dataset with more albums per year and apply the same visualizations to it.

Conclusion/Discussion

During this project the group has made a dashboard and a report describing the visualizations of the dashboard. The dashboard contains visualizations on the dataset "*Best Selling Albums By Duration (1990-2021)".* It attempts to give insight into and answer the four questions in the *project objectives* section. Here, the dashboard shows that there is indeed a correlation between the year and the development of album sales, it answers which genres are the most popular through the years, gives an overview of how many artists are repeated and how many times, and finally it can give recommendations to the reader on which artists and albums they should listen to.

This was done through the use of graphs dedicated to each question, which uses the most effective visual channels for improved readability, and to answer some of the questions both animation and interaction are used. The dashboard is structured with only one page housing all of the visualizations. At first, the dashboard is author-driven as the reader is presented

with the two graphs of *Q1*. After this, they are free to scroll through the page, look at the other graphs, and interact with those that are interactive. This follows the reader-driven format

Regarding improvements to the course, the project group unanimously agrees that there should be more time to work on the project, for example during exercise hours after the day's lecture. The idea here is to have the lectures be two to three hours in length, while the project time then fills up the remainder of the four hours. During this time there should be help available by the teacher and teaching assistants. This proposal would also help hinder the scenario where a group has to work a lot at the end of the semester, as they have not done enough work early on.

References

[1] T. M. Fountain, "Council Post: The Evolution of the Music Industry — and What It Means for Marketing Yourself as a Musician," Forbes, Sep. 13, 2021. https://www.forbes.com/sites/forbesbusinesscouncil/2021/09/13/the-evolution-of-the-music-industry---and-what-it-means-for-marketing-yourself-as-a-musician/?sh=5ad8d2fd297a (accessed Nov. 08, 2022).

[2] N. Adair, "Best Selling Albums By Duration (1990-2021)," www.kaggle.com, Sep. 24, 2022. https://www.kaggle.com/datasets/nickadair44/top-10-annual-best-selling-albums-by-length (accessed Nov. 15, 2022).

[3] personal.sron.nl. (n.d.). Paul Tol's Notes. [online] Available at: https://personal.sron.nl/~pault/#fig:scheme_light [Accessed 8 Dec. 2022].

[4] Wilke, C.O. (n.d.). Fundamentals of Data Visualization. [online] clauswilke.com. Available at: https://clauswilke.com/dataviz/color-pitfalls.html.

[5] S. S. Stevens, "On the psychophysical law.," Psychological Review, vol. 64, no. 3, pp. 153–181, 1957, doi: 10.1037/h0046162.

Contribution Table

Activity	Andreas Edal Pedersen	Benjamin Mikael Andersen	Anni Enevoldsen	Svend Anton Vibæk
Contribution	100%	100%	100%	100%