# Dataviz: An R package to create accessible and cohesive R plots for an edited research book

Jimmy Effendy

7/11/2021

## Contents

# 1 Abstract

The work involved in the productions of visualizations in an edited book is often laborious. With many authors involved, it is difficult to produce an edited book that has a cohesive design. This project aims to reduce workflow required in producing visualizations that are readable, cohesive, and accessible in a Multilevel Regression and Poststratification (MRP) edited book. The project contributes to a ggplot theme and accessible colour schemes. The theme is developed based on Tufte's principle of data graphical excellence. In addition, the project also develops two colour schemes for each type of colour palettes: qualitative, sequential, and diverging palettes. These theme and colour palettes are bundled together in an R package called `dataviz`. An R package is used for this project as authors can easily install the package and use its functions for their research.

# 2 Introduction

This project relates to an edited book about Multilevel Regression and Poststratification (MRP). As is the case with other books, there will be numerous visualizations that will be used within the book. These plots will need to be readable, cohesive, and accessible. There is a group of authors (around one to five authors) who work together on a chapter. The editors then work to combine all the chapters to finalise the book. With many authors contributing to the book, there is a need to balance authors' autonomy so that the final version of the book have a consistent look/design. One mean to achieve this is by producing visualizations that meet a set of criteria.

A style guide can be used to create a cohesive edited book. Implementing a style guide, however, is time consuming and expensive. The editors/designers need to spend a lot of effort and time to update or reproduce figures. Another option that can be taken is by creating an R package that contains theme and colour scale functions that work with `ggplot2` (Wickham, 2016a). This will allow users (in this case, authors) to add a theme and a colour scale function to their plots. This is easy to use, and it will cause minimal disruptions to authors' workflow. More importantly, this will allow the book to have visualizations that are readable, cohesive, and accessible.

There are several aspects that need to be considered to create a cohesive edited book. Firstly, a ggplot theme needs to be developed in a way that allows readers to extract information from the visualizations as much as possible. In addition to theme, various colour scales that are accessible need to be developed for different data types (e.g. categorical and numerical information). The theme and colour scales will be wrapped in an R package called `dataviz`. The R package contains the following items:

- Function for the theme and colour scales.
- Documentations.
- Testing.

In the next section, the project will discuss how a theme that allows viewers to extract information as much as possible from a graph was developed for the project. Next, the project will explore the theory behind colour scales in data visualizations, and describe the methods of choosing colour palettes for the project. In the fourth section, the report will explain how the theme and colour schemes will be wrapped in an R package. To put it briefly, the report contributes themes, colour schemes, and a package called `dataviz`.

# 3 Themes

The first aspect that needs to be considered is how a good visualization looks like, and whether there are general rule that can be applied across different types of plots. However, before looking at principles of good data visualizations, it is better to understand what constitute a visualization. Azzam, Evergreen, and Germuth (2013) stated that data visualization is defined by three criteria:

- A process that is based quantitative or qualitative data
- The result of the process is a graph that is representative of the raw data
- The graph is readable and support exploration, examination, and communications of data.

Without careful consideration, data visualization has the potential to cause confusion resulting in misunderstanding or error in decision making (Evergreen, 2011). In addition, inefficient visual decoding may result in users to not detect important aspect of the data (Cleveland, 1993a). Unfortunately, inefficient graphs are not uncommon occurrence, especially with increasing number of options in data visualization over the years. Tufte (2001) found that prevalent errors and data misrepresentation were found in mass media. Another study found that a total of 1,365 cases of unnecessary graphical elements were found out of 20,080 visualizations from a set of articles published between 1996 and 2016 (Friedman, 2021).

With widespread error in data visualizations, guidelines for better graphs for quantitative and qualitative data were established. According to Tufte (2001), visualizations need to bring viewer's attention to the substance of the data as opposed to graphic design, methodology or something else. In addition, graphs should prevent distortion of the data and should encourage viewers to compare different parts of data. Another suggestion from Tufte is to avoid the usage of chartjunk. Any decorations that do not provide new information to readers are considered to be chartjunk. This can easily cause viewers to be confused or distracted from the data. Some examples of chartjunk are moiré effect (the illusion of vibration caused by excessive patterns), heavy grid lines, noisy background, a fake perspective (i.e. 3-D charts).

Cleveland and McGill (1984) conducted a study of effective data visualization based on graphical perception to investigate humans' ability to visually decode information encoded on graphs. In the study, people were given elementary perceptual tasks to decode information from graphs. The tasks are then ordered based on the accuracy of the performance. The order of the elementary tasks from most to least accurate is:

- Position along a common scale
- Positions along nonaligned scales
- Length, direction, angle
- Area
- Volume, curvature
- Shading, colour saturation

This shows that there are general rules that can be applied across different types of plots, and this can be computationally coded once for efficiency. The next section will discuss how this can be implemented through the use of `ggplot2` (Wickham, 2016a) theme.

## 3.1 Roles of Themes in Data Visualization

Themes in `ggplot` (Wickham, 2016a) package are *non-data* components of data visualizations. Themes allow user to customize plot the background, axis line, ticks, legend position, grid, etc. As a default, `ggplot` (Wickham, 2016a) uses `theme_grey` for plots. However, `ggplot` (Wickham, 2016a) provides other complete themes that users can use, such as `theme_bw()` and `theme_minimal()`.

Theme is especially important for a research book where there are many authors contributing to the book. Without a set theme, it is likely that each author will apply their own visualization styles for their research.

Not only does setting a theme for a research book allow authors to avoid spending their time to customize their plots' appearance, but it also allows research book to have consistency in their visualizations.

As stated in previous section, customizing plot themes is imperative in data visualizations as non-data components impacts readers' ability to comprehend the graph. The next section will discuss how general principles of graphical excellence from Tufte can be applied to themes.

## 3.2 Criteria for good themes

There are five principles of data graphics that are proposed by Tufte (2001). This project focuses on two of the principles that specifically relate to a theme. In his book, Tufte introduced an important concept: data-ink. Data-ink is defined as a non-redundant ink; it is ink used on a graph that represent the underlying data.

$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

Good graphical representations can be achieved by reducing non-data ink as this a way to maximize data-ink ratio (Tufte, 2001). Graphical inks that provide no new information about the underlying data are not beneficial to readers and sometimes result graphical clutters. However, maximizing the data-ink ratio needs to be done cautiously as components of the graph that provide useful information can be removed by mistake.

Design choices on data visualizations need to allow information to be presented in an organised fashion. For example, the colour of non-data components (axis, labels, grid lines, borders) can be grey, and the data component can be brightly coloured, making the data to be the focus of attention (Stone, 2006). This will assist readers to better understand the roles and relationship between different elements of the visualizations.

## 3.3 Potential Themes

The defualt `ggplot2` theme is `theme_grey()`, with grey background and white grid lines. These colour combinations reduces the contrast between the background and grid lines and gives the data plots more value-added appearance (Carr, 1994). Moreover, grey background allows the plot to have a similar colour to the rest of the texts. As a result, the graphics blend well with the flow of a text (Wickham, 2016b).

While having a grey background and white grid lines have their benefits, some steps can be made to improve the theme. As shown in figure 1, having a grey panel background
on top of a white plot background unnecessarily increase non-data components. In addition, having both major and minor grid lines distracts readers from the real data, and rarely needed as data visualizations are not meant to provide precise quantitative value (Few, 2005).

A ggplot theme is developed for this project that is not only based on Tufte principles of data graphics, but also to create a theme that is suitable for a wide range of plots. There are two potential themes that were developed for `dataviz`: `theme_seagull()` and `theme_wombat()`.

### 3.3.1 Seagull Theme

Plot examples with `theme_seagull()` can be seen in figure 2. This `ggplot2` theme is built based on `theme_minimal()`. The idea behind this theme is to reduce non-data components as much as possible by maximizing white space around the plot. The `%+replace%` operator is used to change the elements of the plots. Axis line on the y axis is removed, and axis line on the x axis is thicker. The major vertical grid lines and minor grid lines on both axis are removed. The grid lines, as shown in figure 2, are visually muted so that it does not distract the data points. Cleveland (1993b) found that grid lines enhance readers' perceptions in making visual comparison and accuracy of information extraction.

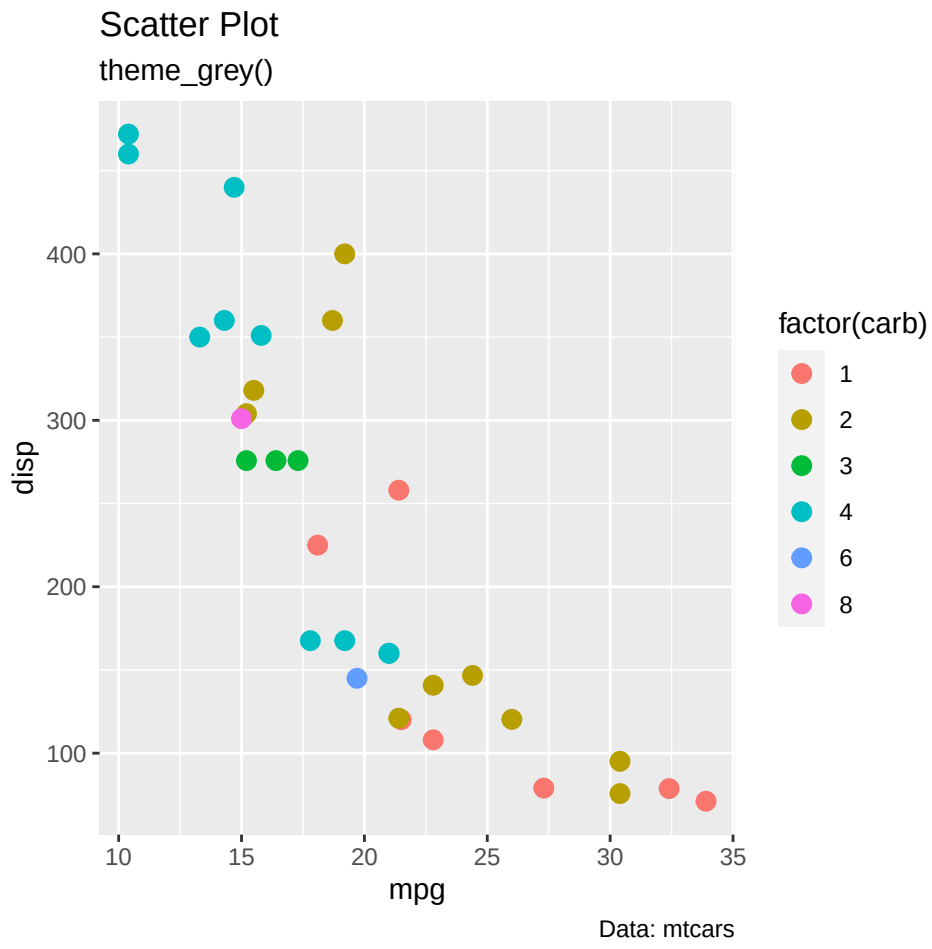



Figure 1: A scatter plot example from `mtcars` dataset with `theme_grey()`. This theme uses light grey background with white grid lines.

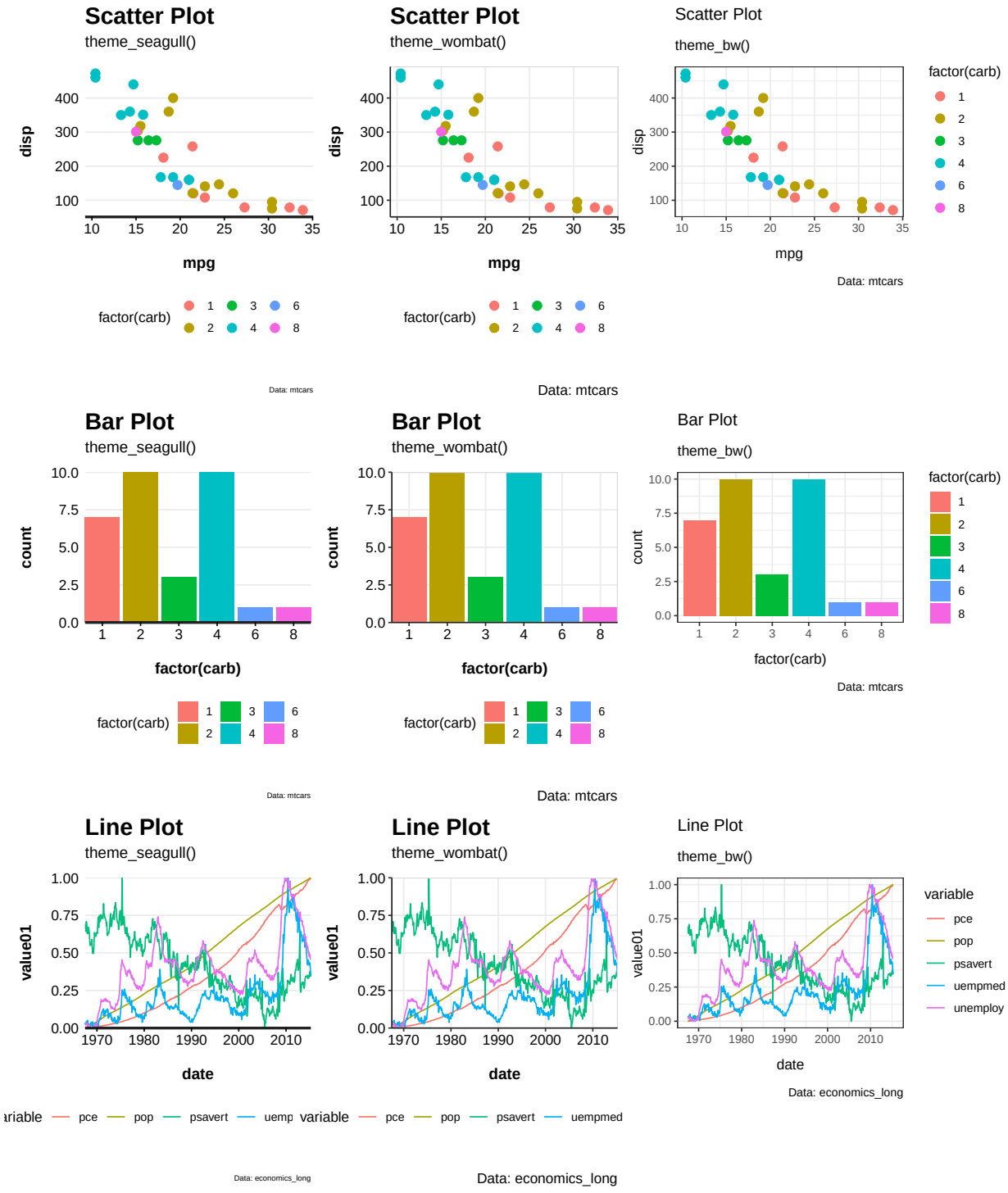



Figure 2: Comparison between `theme_seagull` (left), `theme_wombat` (middle), and `theme_bw` (right). The rows represent different figure types from a data example from R. Basic colour palette from R is used to show categorical variables for this example. In the next section, we will discuss accessible colour palettes that are developed for the project.

### 3.3.2 Wombat Theme

Another potential theme for this project is called `theme_wombat()`. This theme is developed based on `ggplot2` basic theme `theme_bw()`. The `%+replace%` operator is also used to develop `theme_wombat()`. It can be seen in figure 2 that minor grid lines on both axis are removed for this theme. Similar to `theme_seagull()`, `theme_wombat()` has white background and light grey grid lines.

## 3.4 Themes Comparison

This section is going to compare the potential themes `theme_seagull()` and `theme_wombat()` with `ggplot2` (Wickham, 2016a) built in theme, `theme_bw()`. The `theme_bw` is chosen for this comparison as it is one of the built-in themes that minimizes non-data components while still preserving light coloured grid lines.

Side by side comparison of scatter plots, bar charts, and line charts between the three themes can be found in figure 2. One of the differences between `dataviz` potential themes and `theme_bw()` is the size of the text. The axis title, plot title, subtitle, and caption with `theme_seagull()` and `theme_wombat()` are easier to read in comparison to `theme_bw()`. Secondly, the legend position is located on the bottom of the plot as opposed to the right of the plot. The reason to the change in legend position is to maximize the data-ink ratio of the plots. Putting the legend on the right of the plot takes up the space that can be used to display the underlying data. Another way of maximizing data-ink ratio is by reducing the margin around the plot. As shown in the bar plots and line plots in figure 2, there are no extra space between the bars/lines with the axis lines in the plots with `theme_seagull()` and `theme_wombat()`.

One of the objectives of `theme_seagull()` is to have a theme that increases data-ink ratio as much as possible. In comparison to other two themes, `theme_seagull()` has the least amount of grid lines and borders around the plot, making it to have the highest data-ink ratio. While there is no vertical grid lines, major horizontal grid lines are preserved as they aid readers to approximate and compare values between data points.

The `theme_wombat()` is developed to be the middle ground between `theme_seagull()` and `theme_bw()`. In comparison to the `ggplot2` built-in theme, this theme is less cluttered as it does not have minor grid lines on both axis. Unlike `theme_seagull()`, however, this theme retains major vertical grid lines.

### 3.4.1 The Chosen Theme

As stated above, `theme_seagull()` has the least clutter between the two potential themes.However, `theme_wombat()` is likely to assist readers to perform comparison more accurately than `theme_seagull()` while still maintaining a high level of data-ink ratio. Using the scatter plot example in figure 2, the major vertical grid lines from `theme_wombat()` aids readers' ability to estimate the value of data points along the x axis. This shows that `theme_wombat()` is a good compromise of minimizing clutters as well as allowing readers to extract information from the graph as much as possible.

# 4 Choosing Colour Palletes

In addition to themes, colour also plays an important role in data visualizations. Proper use of colour in a visualization enhances the quality of information being presented (Zeileis et al., 2009). One of the most important use of colour in data visualizations is to differentiate one element from one another (Stone, 2006). However, a poor choice of colour palette can impede viewers' ability to effectively extract information from the visualization.

With a simple visual feature, some objects can be easily recognised without requiring much attention. The target object will "pop-out" in field of vision regardless of how many distractors surrounding it (Treisman,

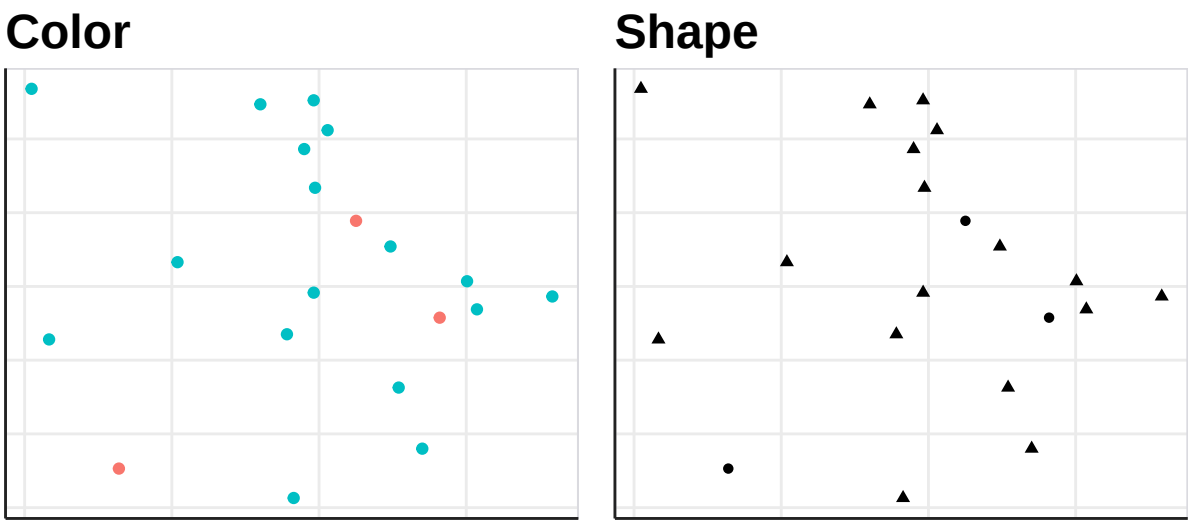



Figure 3: This plot shows the same data with groups highlighted with either colour (left hand panel) or shape (right hand panel). From this, we see that it is easier to find the odd ones by using colour.

1985). The use of colour in data visualization can facilitate this "pop-out" as it makes some elements to be easier to be found.

Figure 3 is an example, which is adapted from Healy's (2018) book about data visualization, that demonstrates how easy it is for colour to be recognised in a visualization. The figure contains three "odd" observations that are different than the rest. To represent this, the first panel uses colour to differentiate the different kinds of observations. The three observations are coloured red, and the rest are blue. On the other hand, the second panel uses shape to represent this feature. Circle is used to represent the "odd" observations, and triangle is used for the rest. While readers have no problem in identifying the three observations in the second panel, it is much easier to find these observations on the first panel. This shows that finding the observations of interest in the first panel requires significantly lower level of attentions compared to the second panel.

In addition to assisting viewers in differentiating one attribute to another, colours can also be used to evoke viewers' emotions. Emotion is important in visualizations for story telling, communicative intent, and engagement (Bartram et al., 2017). It was also found that emotions can significantly improve many cognitive processes such as attention, learning, decision making, and problem-solving (Harrison et al., 2012).

This project will focus to create colour schemes that assist users to easily differentiate one data attribute from another. To achieve this, it is important to choose these schemes using a colour space that works well with human perception of colour. The next section will discuss the different options of colour space, and which one will be used for the project.

## 4.1 Colour Spaces

To effectively use colours in visualizations, it is imperative to understand how human colour perception works. It was theorized that human colour vision evolved in three distinct stages (Ihaka et al., 2003):

1. Perception of light and dark (monochrome)
2. Perception of yellow/blue contrast
3. Perception of green/red contrast

As a result of these three colour axes, colours are naturally described as locations in a three dimensional space. While these colour axis provide physiological description, human perception of colours do not corresponds to these axes. Human perception of colour corresponds to the polar coordinates of the colour coordinate (yellow/blue and green/red), and a dark/light axis (Zeileis & Hornik, 2006).

Unfortunately, not all software packages use colour space was developed based on human perception. For example, the most common colour map used by software packages, the Red-Green-Blue (RGB) (Stauffer et al., 2015), was primarily developed to correspond with colour generation on computer screens rather than to correspond with human perception of colour (Zeileis et al., 2009). Zeileis, Hornik, and Murrell further stated that it is practically impossible for a human to control the perceptual properties of a colour within this colour model as there is no single dimension that relates to hue or brightness of a colour

One popular alternative to RGB system is Hue-Saturation-Value (HSV). This model is a simple transformation of RGB system and aims to capture the dimensions of human visual perceptions (Smith, 1978). Unfortunately, the result of mapping HSV's three dimensions to three dimensions of human perceptions is very poor (Zeileis & Hornik, 2006). Moreover, it is fairly difficult to find a set of colours that are in harmony in HSV colour space (Zeileis & Hornik, 2006). Furthermore, HSV encourages the use of colours that are highly saturated. While these colours are good to attract viewers' attention, it is difficult to look at for a long time (Zeileis et al., 2009).

HCL, which is short for Hue-Chroma-Luminance, was developed to overcome many problems encountered by other colour spaces. This colour space is developed based on the three colour attributes that corresponds to human perceptions: hue, chroma, and luminance (Sarifuddin & Missaoui, 2005). It was found that the three dimensions from this colour space matches human visual perception well (Zeileis et al., 2019). In general, HCL colour space can be used to generate three types of colour scales: qualitative, sequential, and diverging.

#### 4.1.1 Qualitative Palettes

This colour palette is designed for coding categorical information where the categories do not have particular ordering (Zeileis et al., 2019). In other words, this colour palette will give the same perceptual weight to each category so that there will be no perception of one group to be larger or more important than any other one (Zeileis & Hornik, 2006). This colour palette is commonly applied to bar plots or box plots.

#### 4.1.2 Sequential Palettes

Sequential pallets are used for coding ordered/numerical information that span in a certain interval (Zeileis & Hornik, 2006). The low values of the interval are considered to be uninteresting, whereas high values are considered to be interesting (Zeileis et al., 2019). Low values are usually depicted by lighter colours, while darker colours are used for high values (Harrower & Brewer, 2003). This palette is usually used on heatmaps.

#### 4.1.3 Diverging Palletes

Similar to sequential palette, diverging pallets are designed for coding ordered/ numerical information that ranges in an interval. The difference between diverging and sequential palletes is that diverging pallets include a neutral value (Zeileis & Hornik, 2006). This palette should be used when an important data class or break points need to be highlighted (Harrower & Brewer, 2003). Harrower and Brewer further stated that this critical data class or break points is highlighted by a change in in hue and luminance, and should represent a critical value of the data (e.g. mean, median, or zero). For instance, the average GDP per-capita in a choropleth map can be emphasized (i.e. set as neutral value) so that the values above and below the average can be shown in different hues.

## 4.2 Colour palette of choice

In addition to plot themes, this project also contributes two colour schemes for each of the three colour palettes type mentioned in the previous section. The `pals` packages used to build the colour palettes and assess them. The chosen colour palettes need to be assessed to ensure that they achieve their intended purpose. For example, the qualitative colour palette needs to avoid implicit ordering as this palette is intended for representing categorical variables. In addition, we need to ensure that people with colour vision impairments are able to distinguish the different colours in the chosen palettes. The `pal.safe()` function from `pals` package allows us to see how colour palettes are seen by deutan, protan, and tritan colour-blind people.

The methodology of choosing colour palettes that are not only accessible but also enhance the quality of visualizations will be discussed in this section. Firstly, the report will explore common strategies used in choosing qualitative, sequential, and diverging palettes. Then, examples of palettes made from these strategies will be assessed. Lastly, the report will show how colour schemes for qualitative, sequential and diverging palettes are chosen for the project.

### 4.2.1 Chosen Qualitative Palettes

One common strategy to choose this type of palette is by keeping chroma and luminance relatively constant, and hue varied; making all colours in the palette to balance towards the same grey (Zeileis et al., 2009). While this strategy is relatively simple, keeping chroma and luminance fixed may not be the best approach to choose colour palette that are accessible.

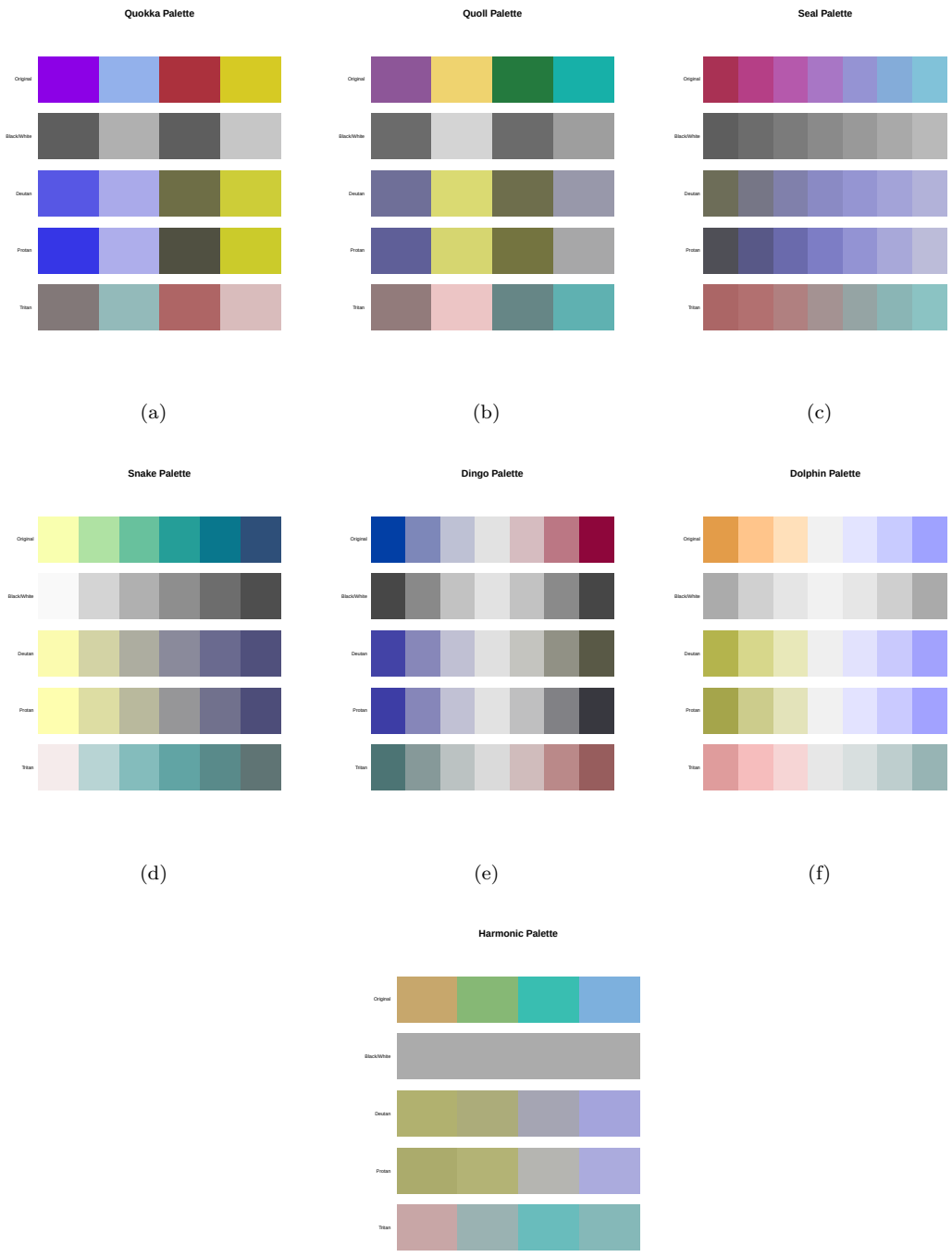
Panel (g) in figure 5 shows an example of a qualitative colour palette from `colorspace` package called harmonic. As shown in the figure, chroma and luminance remain constant, while the hue varies across the four colours. It is fair to say that this colour set works relatively well to show categorical information as it does not give implied ordering across different categories. This method, however, does not have the same result for people with visual impairment. It can be seen in figure 4 that people with colour-blindness will have difficulty in distinguishing some colours from harmonic palette. For instance, there are three colours that look similar as seen by people with tritan colour-blindness. Moreover, this colour palette results in implied ordering of different data group for people with visual impairment.

The two qualitative colour palettes developed for the project are called Quokka and Quoll palettes. Panel (a) and (b) in figure 5 shows the HCL colour spectrum plot for these palettes. In contrast with the Harmonic Palette, Quokka and Quoll palettes have differing hues, chroma and luminance across all colours. As a result, each colour in these palettes can easily be distinguished by people with deutan, protan, and tritan colour-blindness (see figure 4). Furthermore, natural ordering implied by Quokka and Quoll is less prominent than Harmonic palette.

### 4.2.2 Chosen Sequential Palettes

The strategy to choose a sequential palette is different to that of qualitative palette. One easy way to create a sequential palette is by choosing two values for hue (h1 and h2), chroma (c1 and c2), luminance (l1 and l2), and power transformations (p1 and p2) for chroma and luminance (Zeileis & Hornik, 2006). These values of hues, chroma, luminance, and power transformation are used as inputs to `sequential_hcl` function to produce a set of sequential palette. The power transformation allows the chroma and luminance to increase from low values to high values in a nonlinear fashion.

The two sequential palettes created for this project are Seal and Snake palettes. As shown in panel (c) and (d) in figure 5, these palettes use multi-hue schemes as opposed to single-hue schemes. Multi-hue palettes were chosen for the project as it provides better contrast between different values compared to a single-hue sequential palette (Harrower & Brewer, 2003). Moreover, multi-hue sequential palettes avoid the use of extreme colours, as well as allowing viewers to better discriminate the middle colours (Zeileis et al., 2019). In addition, figure 4 shows that people with deutan, protan, and tritan colour blindness are able to




(a) (b) (c)

(d) (e) (f)

(g)

Figure 4: Colour palettes as seen by people with colour vision impairment. The first six colour palettes the palettes that were developed for this project. The last palette, Harmonic palette, is a colour palette from 'pals' package.

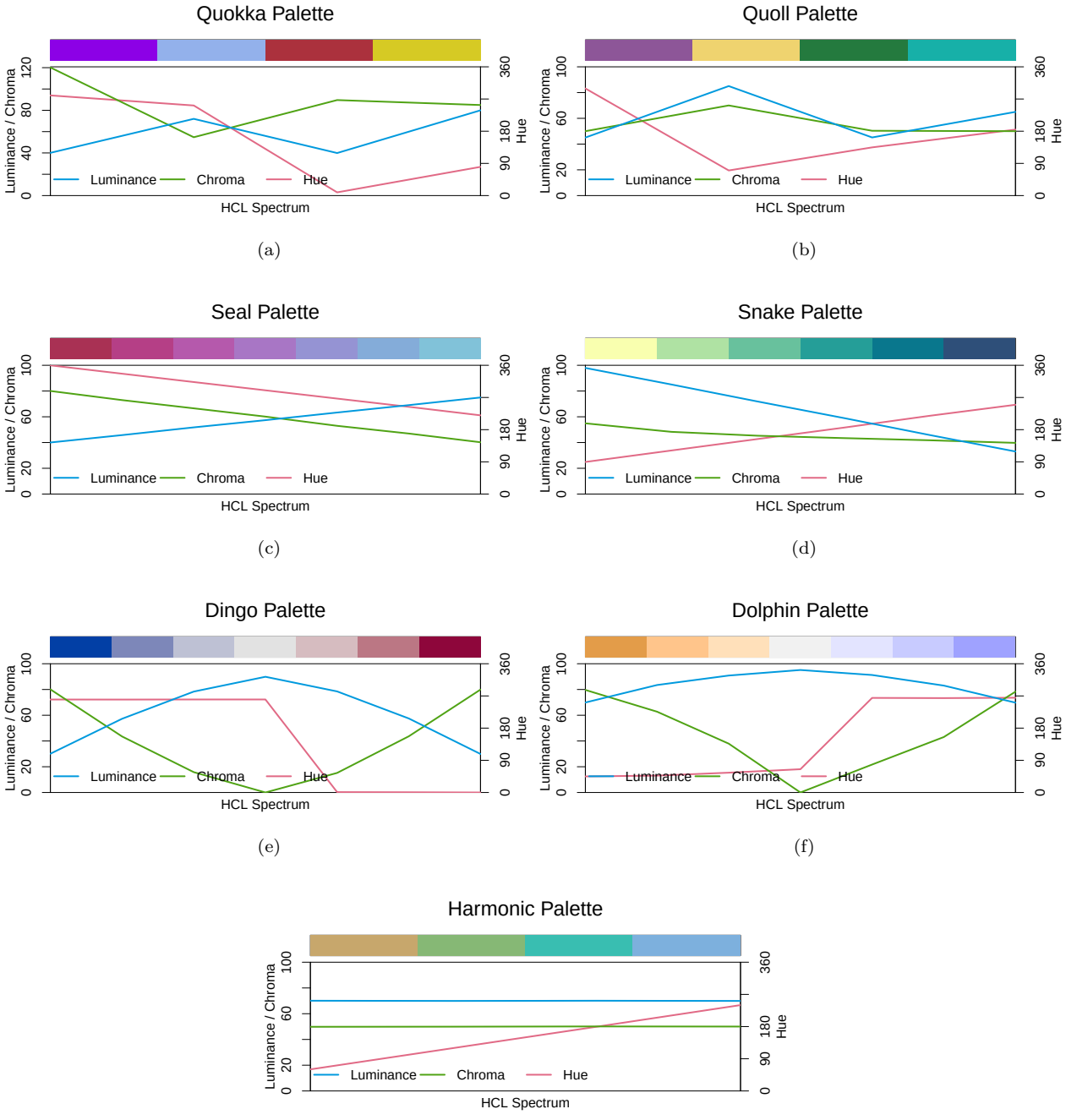



Figure 5: HCL colour spectrum plot for colour palettes of choice. The first six colour palettes the palettes that were developed for this project. The last palette, Harmonic palette, is a colour palette from 'pals' package.

distinguish the colours in the palette. This means that the colour palettes will allow readers with colour vision impairment to see implied order and how the data ranges from low to high values.

#### 4.2.3 Chosen Diverging Palettes

The approach in choosing diverging palettes is similar to choosing sequential palettes. Diverging palettes can be built by choosing two values for hues (h1 and h2), chroma (c1 and c2), luminance (l1 and l2), and power transformation (p1 and p2) for chroma and luminance (Zeileis & Hornik, 2006). These values will be taken as inputs to function `diverge_hcl()` to create a diverging palette. It will automatically choose the neutral colour for the palette.

Three conditions need to be met by a diverging palette to avoid colour confusions. Colour confusions are caused by naming confusions, colour vision impairment, and simultaneous contrast confusion (Brewer, 1996). Naming confusion happens when people use two basic colour names for the same colour (e.g. a colour was called red when presented once, then pink at another time). On the other hand, simultaneous contrast refers to a phenomena where colours are affected by their opposite (e.g. a colour will appear lighter when surrounded by a dark colour, and a colour will appear warmer when surrounded by a cool colour) (Brewer, 1992). This can be prevented by avoiding selecting pairs of colours that are opposite in the colour wheel for diverging palettes (Brewer, 1996).

Dingo and Dolphin palettes are diverging palettes developed for the project. These palettes are shown in panel (e) and (f) in figure 5. Red/blue pairs and orange/blue hue pairs are chosen for Dingo and Dolphin palettes respectively as these pairs create no naming confusions and simultaneous contrast confusions (Brewer, 1996). Moreover, these palettes create no colour confusion due to colour vision impairment as colours in Dingo and Dolphin palettes can be easily distinguished by people with colour-blindness (shown in figure 4). Another reason for choosing these pairs is because red/blue pair is often used in the topic of the edited book, MRP, to show electoral result (e.g. Democratic and Republican party). Therefore, it is important to have a red/blue scale and a non-red/blue scale for functionality in the book.

## 5 R Package

One of the objectives of this project is to create a solution for authors and editors of the research book to create visualizations that are readable, cohesive, and accessible. The solution needs to be easy to use so that it does not disrupt authors' workflow. To achieve this, R code necessary to create the chosen `ggplot2` (Wickham, 2016a) theme and colour schemes will be bundled in an R package called `dataviz`. R packages are extensions to R that can comprise of codes, data, documentation, and tests (Wickham, 2015).

One of the reasons for creating an R package for this project is because of the ease of use. Codes, documentation, and tests in an R package are built in a standardized format so that any R user can install it (Wickham, 2015). In addition, it is fairly simple to create an R package. Packages such as `roxygen2` (Wickham, Danenberg, et al., 2021), `devtools` (Wickham, Hester, et al., 2021), and `testthat` (Wickham, 2011) provide great assistance in documenting and testing the package. Another benefit of using an R package is that any updates to the theme and colour schemes are not time consuming for the users. Any updates to themes and colour schemes simply need to be applied to the code in the R package. There is no need to update the code for each plots in the book; it just needs to be recompiled. This shows that using an R package will allow us to achieve one of the aims of the project: to provide a solution that causes minimal disruption to authors' work.

For users to be able to easily access an R package, it needs to be stored in a repository. For this project, the `dataviz` package will be stored in GitHub. GitHub is a popular choice for a repository as it has the integration with git, a version control software. Version control records changes applied to the package to a file. This allows user to recall specific versions of the package at later time (Loeliger & McCullough, 2012). An alternate option to GitHub is Comprehensive R Archive Network (CRAN). CRAN, however, is not suitable for this package as `dataviz` has a fairly small scope and is intended for a small group of people

(i.e. authors of the MRP research book). CRAN is usually a repository used for more complex R packages intended for large group of people.

In the next section, the report will discuss how the `dataviz` package was created. It will discuss each component of the package, and what tools were used in developing the package.

### 5.1 `dataviz` Components

One of the most important components of a package is the R code. One firm rule for an R package is that its R functions need to be stored in R scripts (Wickham, 2015). Following this rule, the `dataviz` has its R code in R scripts, which are stored in `R` directory. The chosen colour palettes can be accessed from the `scale_colour_dataviz` and `scale_fill_dataviz` functions which are stored in `colour.R`. These functions take the name of the palette as an input (e.g. quokka). The ggplot theme chosen for the project can be used by function `theme_wombat()` which is stored in an R script called `theme.R`.

In addition to R codes, documentation is also a crucial aspect of a well-made package (Wickham, 2015). In the absence of documentation, users will not have a guide that assists them to use the package properly. Proper documentation also assists the authors of the package for future use. This assists authors to remember the purpose of the functions. One type of documentation is an object documentation; this can be accessed by the `?` operator or `help()`. Information such as function description, usage, argument(s), and examples can be found in the documentation. The `roxygen2` (Wickham, Danenberg, et al., 2021), `devtools` (Wickham, Hester, et al., 2021), and `usethis` (Wickham & Bryan, 2021) packages are used in creating object documentation.

Another crucial aspect to package development is testing. While testing add extra work to package development, it ensures that the package works as intended. It also ensures that the package works for edge cases - the less common uses that a user might use. In addition, testing ensures error and warning messages to appear when required (and not appear when they are not supposed to). In other words, testing provide protection and quality assurance for users. We use the packages `testthat` (Wickham, 2011) and `usethis` (Wickham & Bryan, 2021) to write the tests for the `dataviz` package. One example of a test that was used for this package is one that ensures that users use valid input when using `scale_colour_dataviz` and `scale_fill_dataviz`.

## 6 Conclusions and Limitations

One of the challenges in producing a research book is creating a book that has a consistent look or design. Typically, there is a group of authors that work on every chapter of an edited book. These authors will have their own unique style in producing their visualizations for the book. While this is not a bad thing, visualizations style needs to be standardized to keep the look of the book consistent. Not only do these visualizations need to have a consistent design, but they also need to be readable and accessible.

The aim of the project is to provide a solution that can ensure that visualizations in the research book meet the criteria mentioned in the previous paragraph in an efficient manner. To achieve this objective, a ggplot theme was developed for the project. The chosen theme was developed based on Tufte's principles of graphical excellence. In addition, several colour scales for various data types were also created. The colour palettes were developed using a colour space that was based on human perception. Furthermore, to ensure that color is accessible, we visualized how these colours as they would be seen by people with colour visual impairment. These theme and colour scales were developed to ensure that the viewers to extract information as much as possible from the visualizations. The theme and colour schemes are bundled in an R package for efficiency.

One limitation in the package is related to colour palettes, specifically the chosen qualitative palettes. As mentioned in the previous section, this palette is specifically used to show *un-ordered* categorical information. It is therefore imperative to avoid implying that one group is larger or more important than other groups.

However, removing implied ordering while ensuring accessibility proved to be difficult. Figure 4 shows that while the Quokka and Quoll palettes are better than Harmonic palette in avoiding implied ordering for people with colour visual impairment, there are still some colours that look paired. In addition, Quokka and Quoll only have four colours. It is difficult to produce a qualitative palette that have more than four colours while still ensuring that people with colour impairment are able to distinguish the colours from one another.

One area of future work for this project is to work on writing additional testings. For instance, a test can be developed to ensure that a warning message will appear when the theme function is applied to a non-ggplot object. Another test that can be applied to the package is to ensure an error to pop-up when applying a diverging palette to a categorical variable. Another approach that can be done in the future is by having the users (i.e. the authors of the MRP books) to test the functionality of the package for a period of time. This allows the users to provide feedback, and this can be used to improve the package. These testings are beneficial to protect the quality of the package.

In summary, this project contributes an R package containing a ggplot theme and six colour palettes developed to assist MRP book authors in creating visualizations that are readable, cohesive, and accessible.

# Reference


Azzam, T., Evergreen, S., Germuth, A. A., & Kistler, S. J. (2013). Data visualization and evaluation. *New Directions for Evaluation*, *2013*(139), 7–32.

Bartram, L., Patra, A., & Stone, M. (2017). Affective color in visualization. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1364–1374.

Brewer, C. A. (1992). Review of colour terms and simultaneous contrast research for cartography. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *29*(3-4), 20–30.

Brewer, C. A. (1996). Guidelines for selecting colors for diverging schemes on maps. *The Cartographic Journal*, *33*(2), 79–86.

Carr, D. (1994). Using gray in plots. *Statistical Computing & Graphics Newsletter*, *5*(2), 11–14.

Cleveland, W. S. (1993a). A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics*, *2*(4), 323–343.

Cleveland, W. S. (1993b). *Visualizing data*. Hobart Press.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554.

Evergreen, S. D. (2011). Eval+ comm. *New Directions for Evaluation*, *2011*(131), 41–45.

Few, S. (2005). Grid lines in graphs are rarely useful. *Information Management*, *15*(2), 46.

Friedman, A. (2021). Data and visual displays in the journal of ecology 1996–2016. *Information Visualization*, *20*(1), 35–46.

Harrison, L., Chang, R., & Lu, A. (2012). Exploring the impact of emotion on visual judgement. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 227–228.

Harrower, M., & Brewer, C. A. (2003). ColorBrewer. Org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, *40*(1), 27–37.

Healy, K. (2018). *Data visualization: A practical introduction*. Princeton University Press.

Ihaka, R.others. (2003). Colour for presentation graphics. *Proceedings of DSC*, *2*.

Loeliger, J., & McCullough, M. (2012). *Version control with git: Powerful tools and techniques for collaborative software development*. " O'Reilly Media, Inc.".

Sarifuddin, M., & Missaoui, R. (2005). A new perceptually uniform color space with associated color similarity measure for content-based image and video retrieval. *Proc. Of ACM SIGIR 2005 Workshop on Multimedia Information Retrieval (MMIR 2005)*, 1–8.

Smith, A. R. (1978). Color gamut transform pairs. *ACM Siggraph Computer Graphics*, *12*(3), 12–19.

Stauffer, R., Mayr, G. J., Dabernig, M., & Zeileis, A. (2015). Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society*, *96*(2), 203–216.

Stone, M. (2006). Choosing colors for data visualization. *Business Intelligence Network*, *2*.

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, *31*(2), 156–177.

Tufte, E. (2001). *The visual display of quantitative information*. Cheshire: Graphic Press.–2001.–213 p.

Wickham, H. (2011). Testthat: Get started with testing. *The R Journal*, *3*, 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

Wickham, H. (2015). *R packages: Organize, test, document, and share your code*. " O'Reilly Media, Inc.".

Wickham, H. (2016b). *ggplot2: Elegant graphics for data analysis*. Springer New York.

Wickham, H. (2016a). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., & Bryan, J. (2021). *Usethis: Automate package and project setup*. https://CRAN.R-project.org/package=usethis

Wickham, H., Danenberg, P., Csárdi, G., & Eugster, M. (2021). *roxygen2: In-line documentation for r*. https://CRAN.R-project.org/package=roxygen2

Wickham, H., Hester, J., & Chang, W. (2021). *Devtools: Tools to make developing r packages easier*. https://CRAN.R-project.org/package=devtools

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2019). Colorspace: A toolbox for manipulating and assessing colors and palettes. *arXiv Preprint*

*arXiv:1903.06490*.
Zeileis, A., & Hornik, K. (2006). *Choosing color palettes for statistical graphics*.
Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, *53*(9), 3259–3270.