

# animbook: Visualizing changes in performance measures and demographic affiliations using animation

by Krisanat Anukarnsakulchularp

**Abstract** An abstract of less than 150 words.

## 1 Introduction

The concept of “zombie companies” began to attract attention when an article on the proliferation of zombie companies (Caballero, Hoshi, and Kashyap (2008)). The “zombie companies” are generally defined as companies with an interest coverage ratio of less than one for a period of more than three years. However, there is a simpler and more easily understandable way to show these concepts by visualizing the new listings (enters) and the de-listing (exits) of publicly traded companies on a country-by-country basis. This visualization method makes it clear that the concepts of zombie companies are not unique to Japan, as indicated in the OECD report (McGowan, Andrews, and Millot (2017)) that the United States has a faster metabolize (more new listings and exits) relative to Japan.

The visualization above can also be thought of as a movement between groups, which is how many companies have entered the market and how many have exited the market. One example of this visualization is from The New York Times article “Extensive data shows punishing reach of racism for black boys” (Badger et al. (2018)). The animation portrays the trajectory of how boys with different demographics could land up in society. It can be adapted to our problems by capturing the movement between the groups over a specific period.

With an advancement in technology, this results in an increase in both the size and complexity of the data. It requires an experienced hand to convey the right messages from the data. However, as shown in The New York Times article (Badger et al. (2018)), the animation is not only created just for eye-catching graphics. Animation can be used as a tool that helps communicate complex data, enhancing the narrative and keeping it engaged for the audience. Based on Mayer and Moreno (2002), animation can improve learning, especially when the goal is to promote deep understanding. It also requires the designer to understand how people learn. The cognitive theory of multimedia learning (Mayer (2005)) will be considered during the designing period.

The challenge of the New York Times article animation was the amount of code needed to reproduce. Additionally, the business data may or may not have the same information as the one in the article. These two purposes establish the objective for creating the R package that could generalize the animation to a wider range of data.

The structure of this paper will consist of both the visualization design and software design. The first section will explain about the animation in The New York Times article. Secondly, what are the expected data to be input into the plot, and how it is processed. The third section is about the animation tools, for example, **gganimate** (Pedersen and Robinson (2020)) and **plotly** (Sievert (2020)). The next section will be about the design of the package and visualization. The last two sections explore the usage of the R package and the application.

## 2 Explanation of the NY Times visualization

## 3 Data

When working with the raw data, the data pre-processing is one of the important steps we need to take before further analysis is taken. The function in this package will offer some flexibility in the types of the variable it accepted. However, there are still some restrictions the user needs to follow. In this section, it will provide what are the accepted format and examples for the user the followed to reproduce the final animate plot.

The data set that will be used in this section and in all of the examples of this journal is collected from Bureau van Dijk. The data set contained 30,000 rows and 94 variable of information on listed, and major unlisted/delisted companies across the world. From the raw data, we are only interested in the ranking of the companies. The cleaned version of this data set is included in the package which only contained the relevant variables to reproduce the animated plot.

## 4 Animation tools

Literature review of animation tools

## 5 Visualization design

This explains how to get from data to the animations, including different sorts of plots.

In designing the package for reproducing the New York Times animated plot, the package ended up with a three-step process in recreating the animation. The first step is to turn the data into the right format for the plot function. The next stage is to create a ggplot object, which can then be inputted into the animation function. The last step is adding the animation settings to the ggplot object so the user can animate the plot using the `gganimate::animate()` function. The reason for this three-step process is that it allows the user who does not have a lot of experience to reproduce the animation while keeping the customization for an experienced user.

As per motivations, the main focus for this package was to look at movement between the percentile group. Thus, the `anim_prep()` function's first purpose was to be able to take any numerical values and output the percentile group of the observations. The algorithms behind the group splitting are the `quantile()` and `cut()` functions. The `quantile` function from the stats R packages (R Core Team (2013)) takes in the numeric vector and outputs the corresponding quantiles to the given probabilities. The output from the `quantile` function can then be used as the `breaks` argument for the `cut` function that is from the base R packages. The concern of this method is that the users might not always be interested in ranking the observations.

The next logical design is to allow for different scaling. As mentioned in the paragraph above, the first scaling for the function is "rank". The function then expanded to allow for the "absolute" scaling. This scaling calculates the quantile groups based on the absolute values scales. The default is to break the group equally using the `seq()` function. The input to the `seq` function is taken by the min and max of the values and increments by equal steps depending on the number of groups of interests. In the real-world scenario, it is not always the case to break the group equally. The users may be interested in the groups that are not equally broken. The function then allows the user to provide the vector of the breakpoint.

## 6 Software

### Installation

### Overview of functions

There are three main stages for this packages, one is to prepare the data, two is plotting the data, and three, animating the plot.

### Prepare the data

The first step can be done using the `anim_prep()` function. This function required that the users data contained `id`, which is used to uniquely identify each individual observation, `values`, which is numerical values to be used to grouping the observation together, and `time`, the variable that associated the changes of the observation. In the cases when the users already have the values as a group variable, the `anim_prep_cat()` function can then be used instead.

The additional options for the `anim_prep()` that allow for more customization to the data structure or the plot are as follows:

- `label`: group labeling.
- `ngroup`: number of groups we want to split the values into.
- `breaks`: the group bins size
- `group_scaling`: the grouping variable for the bins calculations.
- `color`: the variable used to color the observations.
- `time_dependent`: logical. Whether we want the observations to start at the same time or not.
- `scaling`: the scaling method, either rank or absolute.
- `runif_min`: minimum value for random addition to frame numbers
- `runif_max`: maximum value for random addition to frame numbers

Using the different combination of the additional options, the users could end up with four different scaling.

1. Rank scaling
2. Absolute scaling
3. Rank scaling by group
4. Absolute scaling by group

For the `anim_prep_cat()` function, the additional options are:

- `label`: group labeling.
- `order`: the ordering of the group.
- `color`: the variable used to color the observations.
- `time_dependent`: logical. Whether we want the observations to start at the same time or not.
- `runif_min`: minimum value for random addition to frame numbers
- `runif_max`: maximum value for random addition to frame numbers

Both the `anim_prep()` and `anim_prep_cat()` function will return “animbook” object containing a list of the standard format data and settings.

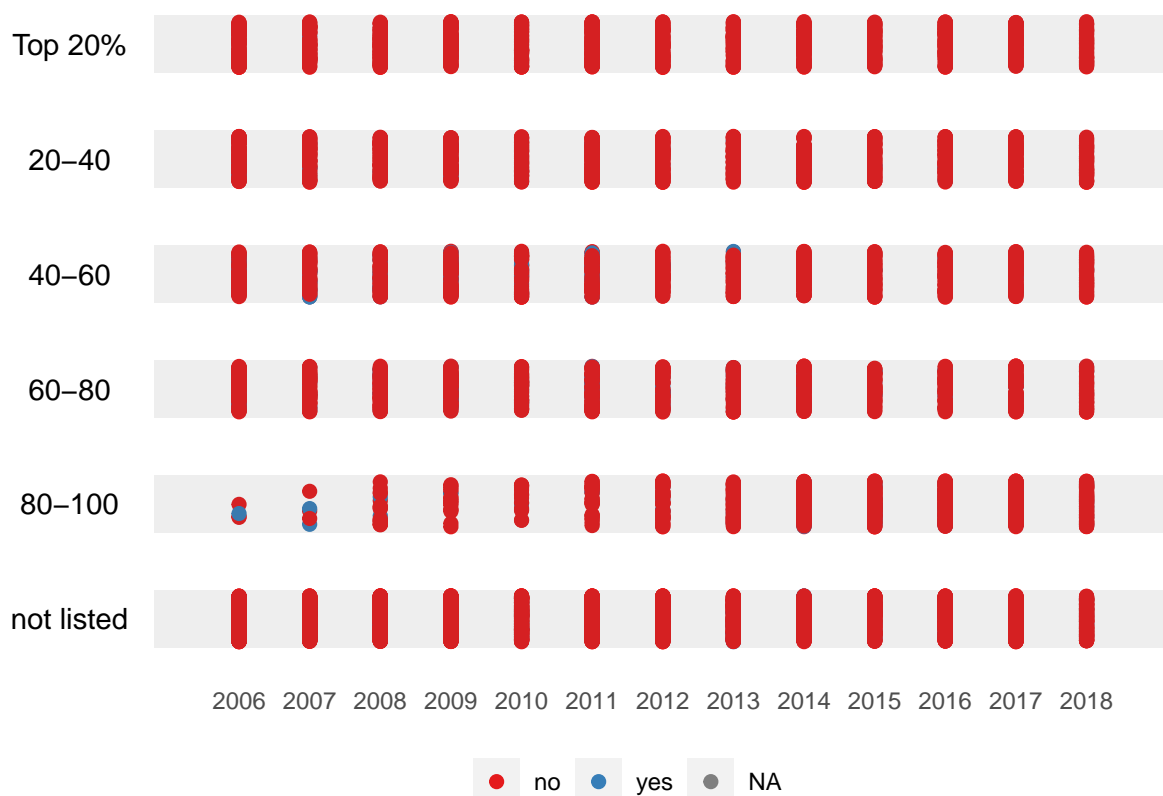
### Plotting the data

Once the data is prepared, the next steps is to created the `ggplot` object as a basis for the animation. There are three plots available in this package, two plot could be uses for the animation and other plot is used as a static visualization.

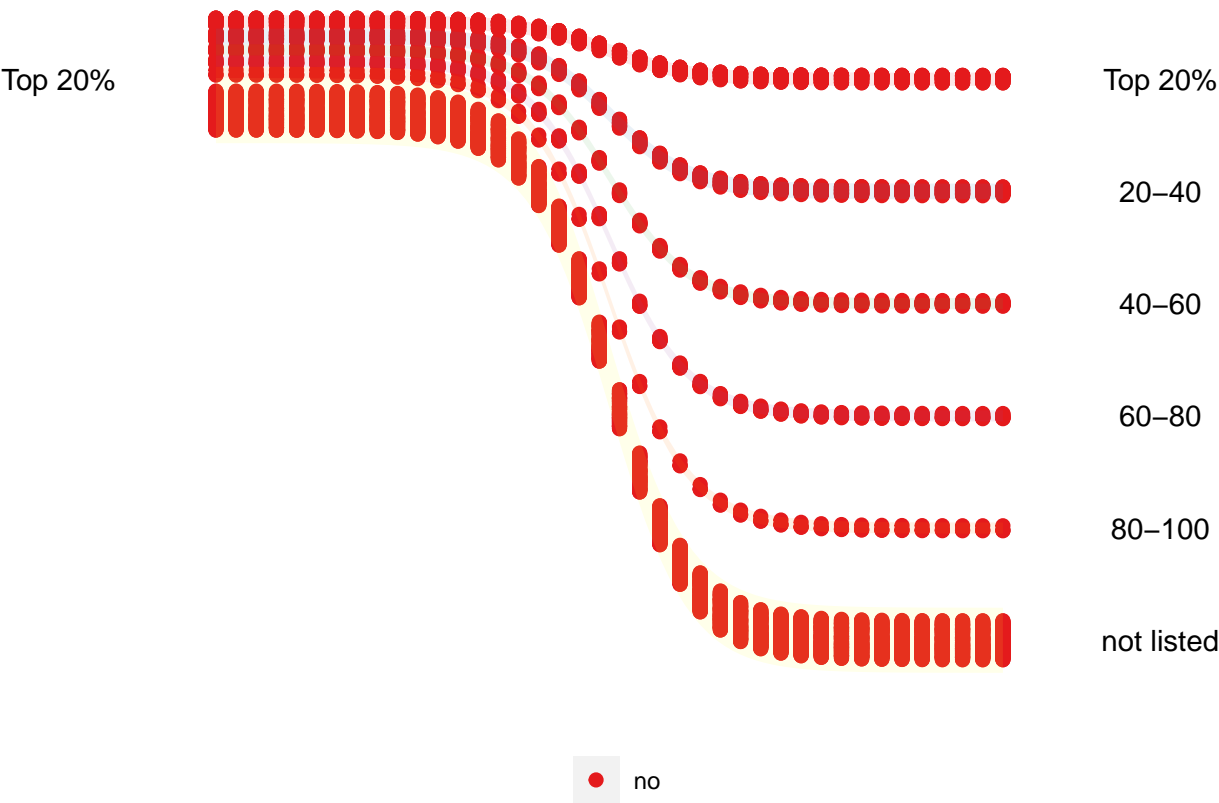
- `kangaroo_plot()`, plots the observation’s movement over time.
- `wallaby_plot()`, the subset plot of the `kangaroo_plot` with the time limit to only start and end.
- `funnel_web_plot()`, the faceted static plot by time variable.

All of the plot, have an internal function that converted the standard data format into the required structure for each plotting function.

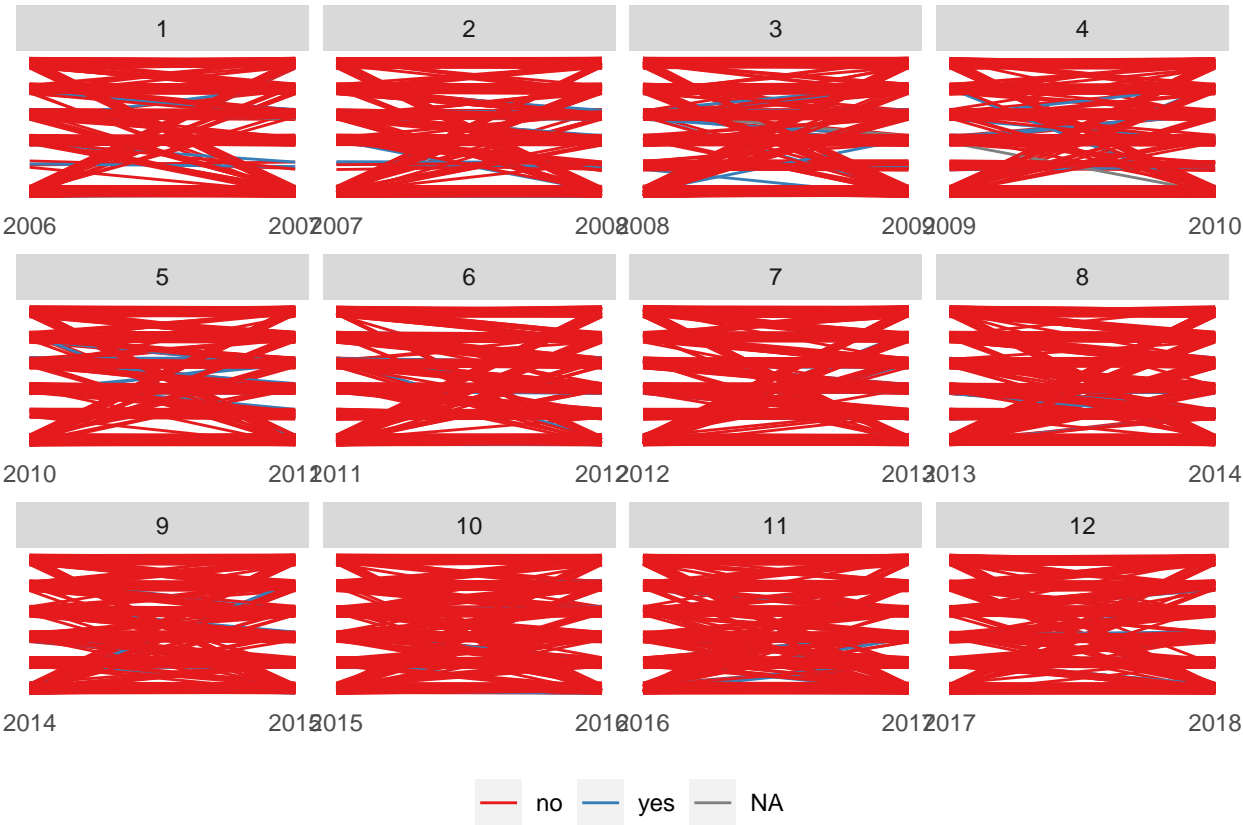
1. Kangaroo’s plot



2. Wallaby’s plot



3. Funnel web spider’s plot



## Animating the plot

To animate the plot, the plot need to be save as an object before passed on to the final function `anim_animate()`.

## Example usage

## 7 Application

### Accounting database: osiris

### Voter behavior

Based on the 2016 Australian election results, how does the top party perform in keeping the old voters for different genders.

## References

- Badger, Emily, Claire Cain Miller, Adam Pearce, and Kevin Quealy. 2018. "Extensive Data Shows Punishing Reach of Racism for Black Boys." *The New York Times*. The New York Times. <https://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html>.
- Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap. 2008. "Zombie Lending and Depressed Restructuring in Japan." *The American Economic Review* 98 (5): 1943–77. <http://www.jstor.org/stable/29730158>.
- Mayer, Richard E. 2005. "Cognitive Theory of Multimedia Learning." In *The Cambridge Handbook of Multimedia Learning*, edited by Richard Editor Mayer, 31–48. Cambridge Handbooks in Psychology. Cambridge University Press. <https://doi.org/10.1017/CB09780511816819.004>.
- Mayer, Richard E., and Roxana Moreno. 2002. *Educational Psychology Review* 14 (1): 87–99. <https://doi.org/10.1023/a:1013184611077>.
- McGowan, Müge Adalet, Dan Andrews, and Valentine Millot. 2017. "The Walking Dead?" no. 1372. <https://doi.org/https://doi.org/https://doi.org/10.1787/180d80ad-en>.
- Pedersen, Thomas Lin, and David Robinson. 2020. "Gganimate: A Grammar of Animated Graphics." *R Package Version 1* (7): 403–8.
- R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Sievert, Carson. 2020. "Interactive Web-Based Data Visualization with r, Plotly, and Shiny." Chapman; Hall/CRC. <https://plotly-r.com>.

Krisanat Anukarnsakulchularp  
 Monash University  
 Faculty of Business and Economics  
 Melbourne, Australia  
 ORCID: 0009-0008-5638-7124  
[kanu0003@student.monash.edu](mailto:kanu0003@student.monash.edu)