# The Unsinkable Titanic Data

Jürgen Symanzik * Michael Friendly † Ortac Onder ‡

**Abstract**
Many readers are likely familiar with the stories of the tragic fate of passengers and crew of the *RMS Titanic* upon her fatal collision with an iceberg and her sinking in the early hours of April 15, 1912, on her maiden voyage to New York City. Little known is the fact that the first graphical summary of the initial survivor data appeared in The Sphere, a British newspaper, on May 4, 1912. The public inquiries that followed produced detailed data sets that have been widely used to illustrate graphical and statistical methods for quite some time. Numerous follow-up studies have used a wide variety of graphical representations related to the *Titanic* disaster, published in statistics, information visualization, and social sciences venues. It seemed timely to survey the variety of graphical methods used for these data sets over the last century. Graph types used to portray the *Titanic* data include: dot plots, bar charts, mosaic plots, doubledecker plots, parallel set plots, Venn diagrams, balloon plots, nomograms, and tree diagrams to name only a few. In this article, we provide an overview of variants of the *Titanic* data set and resulting visualizations.

**Key Words:** Data Visualization; Graphs; Categorical Data Analysis; *RMS Titanic*

## 1. Introduction

We recently discovered a remarkable and little known graph by G. Bron portraying the deaths among passengers and crew on the maiden (and only) voyage of the *RMS Titanic* (Figure 1). This graph (Bron, 1912, p. 103) was published in the London illustrated newspaper, *The Sphere*, on May 4, 1912, less than one month after the ship sank in the early hours of April 15, 1912, upon her fatal collision with an iceberg. About 1,500 passengers and crew, out of about 2,200 aboard the ship, were killed. While the sinking of the *Titanic* is not the largest maritime disaster with respect to the number of lives lost, it is one of the most memorable ones because of the detailed data on lives lost (and not-lost) that have been made available and its huge influence on pop culture via books, movies, TV documentaries, and *Titanic*-inspired exhibits and museums.

Before we go into details on data and graphs related to the fate of the *Titanic's* passengers and crew, we want to start with a brief summary that lead to this disastrous event. The design for the *Titanic* was approved in July 1908 and construction began in March 1909. About three years later, her maiden voyage started in Southampton on April 10, 1912. Only four days later, on April 14, 1912, at 11:40pm, an iceberg struck the *Titanic* on the starboard (right) side. On April 15, 1912, 2:05am, the last lifeboat left the *Titanic* with over 1,500 people still left on the ship. After breaking apart, the last major part, the stern, of the *Titanic* sank on April 15, 1912, 2:20am, less than three hours after the collision, killing about 1,500 out of about 2,200 passengers and crew. Further historical details can be found at `https://www.historyonthenet.com/titanic-timeline-3/` or in any of the non-fiction books that deal with the *Titanic* disaster.

*Department of Mathematics and Statistics, Utah State University, Logan, UT 84322–3900, USA. E-mail: symanzik@math.usu.edu

†York University, Psychology Department, Toronto ONT M3J 1P3, Canada. E-mail: friendly@yorku.ca

‡York University, Schulich School of Business, Toronto ONT M3J 1P3, Canada. E-mail: oonder16@schulich.yorku.ca

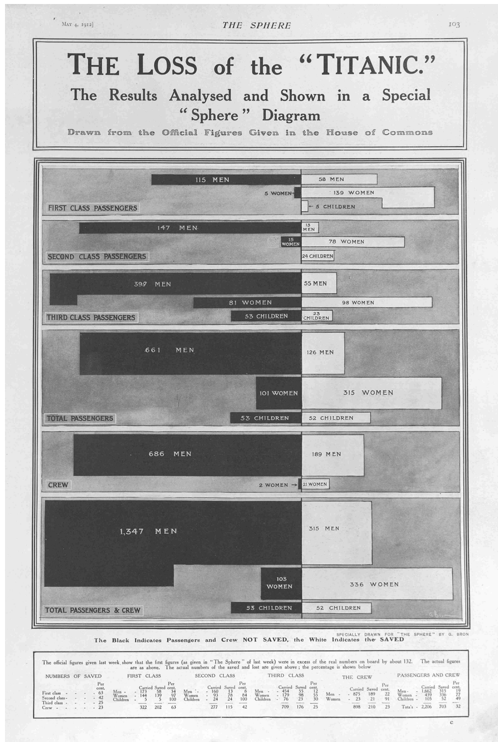



**Figure 1**: The Loss of the "Titanic", specially drawn for "The Sphere" by G. Bron. Originally published in "The Sphere," p. 103, May 4, 1912.

After discovering G. Bron's graph, we decided to track down and catalog the wide variety of graphs and statistical methods used to display and analyze the data related to this disaster. This full-length article is an extension of our overview article in *Significance* (Friendly et al., 2019) and is based on two of our recent conference presentations on this topic, one given at *CompStat 2018, Iasi, Romania* (Symanzik et al., 2018) and another one given at the *Joint Statistical Meetings 2019, Denver, Colorado, United States* (Symanzik et al., 2019). In Section 2 of this article, we provide a brief overview on G. Bron, the creator of the first graph that is based on the *Titanic* data. We introduce the main data sources and data sets related to the *Titanic* disaster in Section 3. Section 4 provides a detailed overview of modern graphs and uses of these data, primarily within statistics, computer science, and social sciences. We present the Info Vis approach to these data in Section 5 and provide a brief overview of competitions that made use of these data in Section 6. We finish with a short discussion in Section 7.

An accompanying web page at `http://www.datavis.ca/papers/titanic/` has been constructed as a supplement to our presentations and articles. While the latter ones explained the context in more detail, there was insufficient space for all illustrations that might be of interest. More importantly, our web page collects the wide variety of images, sources, and references we have found dealing with the *Titanic* data. New materials related to G. Bron, new *Titanic* data sets, and additional sources of graphs inspired by the *Titanic* data will also be added to this web page when they become available.

## 2. G. Bron

G. Bron was a prolific technical illustrator who worked for *The Sphere* and other similar publications about 1910–1925. Little about him is known. There is even confusion regarding his first name.

As far as we can determine, G. Bron was the pseudonym used by a graphic artist, illustrator and cartoonist named either George Treeby or William Brown Treeby. The only reliable information we have found connecting "G Bron" with these other names comes from Australian sources. His biography at `https://www.daao.org.au/bio/george-treeby/`, Design and Art Australia, describes him as a *"Federation-era Melbourne magazine cartoonist, illustrator and writer. Treeby contributed drawings to the* Bulletin *and* Melbourne Punch *as well as contributing articles to* Lone Hand*."*

In its issue from January 14, 1909, p. 27, the *Trove*, published an article titled *"G. Bron (W. B. Treeby)"*. This article provides evidence for the name **William Brown** Treeby. Some quotations from this article article follow:

> Asked to talk about himself, G. Bron remarked that he was 'William Brown Treeby (for age see photograph), father of Sid Treeby, Mab Treeby, Ethel Treeby (R. A. Kent), and a few others.' Further: I was born in London, and brought to Australia by adventurous parents in infancy. Shirking real graft, I suffered an apprenticeship to the unreasonable occupation of wood-engraving, which was wearing to the eyes but good for the patience. [...] Tripped it to England (I, not Job) with wife and young family, and stopped there five years, it being while working in London, half engraver, half artist, that the name 'G. Bron' was invented — a duplication that I sometimes would like to send to pot.

Some evidence for **George** Treeby comes from the Personal section of *The Feilding Star*, Volume XI, Issue 2391, July 10, 1914:

> Victoria has become notable for the number of its literary and artistic, families — the Lindseys, the Dysons, the McCraes, the O'Farrells, and the Treebys. The last-named family, which is now settled in London, consists of three artists and one writer, all of whom are well known in Australia. Mr George Treeby, the father, who draws for the Sydney Bulletin under the name of "G. Bron," has contributed to the Illustrated London News and Graphic, but now devotes all his time to The Sphere. [...]

Some of G. Bron's less known graphs and illustrations have been collected on our accompanying web page at `http://www.datavis.ca/papers/titanic/`. His most important graph from a visualization perspective, *The Loss of the "Titanic"* graph shown in Figure 1, did not only catch our attention. It was also cited, mentioned, or reprinted in several other sources in the statistical and the Info Vis communities over the past ten years, e.g., in Rendgen and Wiedemann, J. (Ed.) (2012), p. 75, Harrell (2015), p. 291, and Feldman (2018), p. 55.

## 3. Data Sources

In this section, we provide references to various sources and data sets on the *Titanic*, including an overview of packages developed for R, a language and environment for statistical computing (R Core Team, 2019), that contain one out of the multiple existing versions of these data.

### 3.1 Primary Sources

Two major inquiries on the sinking of the *Titanic* were conducted in 1912. One of them, the *British Board of Trade Inquiry*, often referred to as *"Lord Mersey Report"*, was conducted by the *British Board of Trade* under the *Right Hon. Lord Mersey, Wreck Commissioner of the United Kingdom*. Hearings took place on thirty-six days between May 2, 1912, and July 3, 1912. Numerous details on passengers and crew aboard the *Titanic* can be found in this report, however in an unorganized way.

The other one, the *U.S. Senate Inquiry*, was conducted by a Subcommittee of the Committee on Commerce, United States Senate, New York, N. Y. Hearings took place on eighteen days between April 19, 1912, and May 25, 1912. Similar to the *British Board of Trade Inquiry*, numerous details on passengers and crew aboard the *Titanic* can be found here. Moreover, this report provided detailed information on the use and release of the lifeboats of the *Titanic*.

Electronic transcripts of the British and U.S. Senate inquiry reports are accessible at `https://www.titanicinquiry.org/downloads/BritishInquiry.pdf` and `https://www.titanicinquiry.org/downloads/USInq.pdf`, respectively. In particular, the web page at `https://www.titanicinquiry.org/` was set up to provide interested readers and researchers access to the original inquiries and other *Titanic*-related information that is otherwise hard to obtain.

### 3.2 Online Data Collections

Several online data sources exist on the web. Most notable is the *Encyclopedia Titanica* web site at `https://www.encyclopedia-titanica.org/`. This site was started in 1996 as an attempt to tell the story of every person that traveled the *Titanic* as a passenger or crew member. It contains numerous interactive lists with full details such as full name, age, class/dept, ticket, joined, job, survived?, boat/body, URL, and photo (if available). Their starting point was the passenger list compiled by Michael A. Findlay for the book *Titanic Triumph and Tragedy* (Eaton and Haas, 1986). This list can be found in other sources as well, e.g., in Geller (1998).

The *ICYouSee* web site at `http://icyousee.org/` was created and is maintained by John R. Henderson. It went online in December 1994. The *Titanic* page of this site at `http://www.icyousee.org/titanic.html`, was first released on June 6, 1998. It provides demographics of the *Titanic* passengers, such as deaths, survivals, nationality, and lifeboat occupancy.

Rebecca Bilbro provided a version of the *Titanic* data set at `https://www.kaggle.com/c/titanic/data` for use in the Kaggle Competition *"Predicting Survival Aboard the Titanic,"* described in more detail in Section 6. The same data can also be found in the R package `titanic`, listed in the overview in the next section.

### 3.3 R Data Sets

The *Titanic* data first appeared as a real data set (cases by variables) in connection with a *Journal of Statistics Education* article (Dawson, 1995). These data and their description are directly accessible at `http://jse.amstat.org/datasets/titanic.dat.txt` and `http://jse.amstat.org/datasets/titanic.txt`, respectively.

A variety of other forms and versions of these data are available in R and R packages now. Therefore, when referring to the *Titanic* data set, it is essential to indicate the exact source in R and ideally the original data source used for the creation of this R data set.

There are differences in the number of observations, the number of variables, and even what the actual count of lives lost and lives saved is, based on the original historical source.

As of September 2019, there exist at least 12 different R packages with a total of 17 different *Titanic* data sets. In the following overview, these data sets are given in the form `package::dataset`: Title — Description (Format). In R, use `?package::dataset` for a more detailed description of the contents and original sources of these R data sets.

- `carData::TitanicSurvival`: ‘’Survival of Passengers on the Titanic” — Information on the survival status, sex, age, and passenger class of 1309 passengers in the Titanic disaster of 1912 (a data frame with 1309 observations of 4 variables).
- `COUNT::titanic`: “titanic” — The data is an observation-based version of the 1912 Titanic passenger survival log (a data frame with 1316 observations of 4 variables).
- `COUNT::titanicgrp`: “titanicgrp” — The data is a grouped version of the 1912 Titanic passenger survival log (a data frame with 12 observations of 5 variables).
- `DALEX::titanic`: “Passengers and Crew on the RMS Titanic Data” — The titanic data is a complete list of passengers and crew members on the RMS Titanic. It includes a variable indicating whether a person did survive the sinking of the RMS Titanic on April 15, 1912 (a data frame with 2207 rows and 9 columns).
- `datasets::Titanic`: “Survival of passengers on the Titanic” — This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, summarized according to economic status (class), sex, age and survival (a 4-dimensional array resulting from cross-tabulating 2201 observations on 4 variables).
- `earth::etitanic`: “Titanic data with incomplete cases removed” — Titanic data with incomplete cases, passenger names, and other details removed (a data frame with 1046 observations on 6 variables).
- `msme::titanic`: “Titanic passenger survival data” — Passenger survival data from 1912 Titanic shipping accident (a data frame with 1316 observations on 4 variables).
- `PASWR::titanic3`: “Titanic Survival Status” — The titanic3 data frame describes the survival status of individual passengers on the Titanic. The titanic3 data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers (a data frame with 1309 observations on 14 variables).
- `rpart.plot::ptitanic`: “Titanic data with passenger names and other details removed” — Titanic data with passenger names and other details removed (a data frame with 1046 observations on 6 variables).
- `stablelearner::titanic` : “Passengers and Crew on the RMS Titanic” — The Titanic data is a complete list of passengers and crew members on the RMS Titanic. It includes a variable indicating whether a person did survive the sinking of the RMS Titanic on April 15, 1912 (a data frame containing 2207 observations on 11 variables).
- `Stat2Data::Titanic`: “Passengers on the Titanic” — List and outcomes for passengers on the Titanic (a data set with 1313 observations on 6 variables).

- `titanic::titanic`: "titanic: Titanic Passenger Survival Data Set" — titanic: Titanic Passenger Survival Data Set.
- `titanic::titanic_gender_class_model`: "Titanic gender class model data" — Titanic gender class model data (a data frame with 2 columns).
- `titanic::titanic_gender_model`: "Titanic gender model data" — Titanic gender model data (a data frame with 2 columns).
- `titanic::titanic_test`: "Titanic test data" — Titanic test data (a data frame with 11 columns).
- `titanic::titanic_train` : "Titanic train data" — Titanic train data (a data frame with 12 columns).
- `vcd::Lifeboats`: "Lifeboats on the Titanic" — Data from Mersey (1912) about the 18 (out of 20) lifeboats launched before the sinking of the S. S. Titanic (a data frame with 18 observations and 8 variables.).

The `datasets::Titanic` version of the *Titanic* data set is the first one that was released in R. It is based on the data from Dawson (1995) and was released in R, version 0.90.1 in December 1999. The `NEWS` that accompanied this version of R indicated the following as one of the new features of this version: *"New data sets 'HairEyeColor' (hair and eye color of statistics students), 'Titanic' (survival of passengers on the Titanic), and 'UCBAdmissions' (student admissions at UC Berkeley)."* It is notable that three well-known statistical data sets all were added to R at the same time.

Three different versions of the *Titanic* data set are available from the data sets archive at the Department of Biostatistics at Vanderbuilt University at `http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`. The web page at `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html`, created by Frank E. Harrell, Jr, provides details on the content and sources for each data set. The `titanic` data set is available in R, S-Plus, and ASCII format. The `titanic2` data set is available in R and S-Plus format. The `titanic3` data set is available in R, S-Plus, Excel, and ASCII format. These three data sets consist of 1,313 observations and 10 variables (`titanic`), 2,201 observations and 4 variables (`titanic2`), and 1,309 observations and 14 variables (`titanic3`). This last version is also directly available in R from the `PASWR` R package.

### 3.4 Variations

Dawson (1995), paragraph 7, noted:

> More detailed research into the Titanic disaster revealed some differences of opinion on the number lost. For instance, the Encyclopaedia Americana (1994) gives the death toll as "variously estimated as 1,490, 1,502, and 1,517." A book edited in 1912 under the pseudonym Marshall Everett gives the figure variously as 1635 and 1595 (Everett 1912); the first of these figures agrees with that found in Logan Marshall's book (Marshall 1912). However, the British Board of Trade Inquiry Report (1990), written originally in 1912, claims a death toll of 1490. Modern sources seem to agree that the true numbers are in the neighborhood of 1,500, but the exact numbers may never be known.

In the end, Dawson's data set (and thus the first `datasets::Titanic` data set used in R) consisted of 2,201 observations: 1,316 passengers and 885 crew. From these, 711 survived and 1,490 did not survive.

In contrast, the table underneath G. Bron's chart in Figure 1 lists the following numbers that are plotted in the bars. Based on G. Bron's data, there were 1,318 passengers and 898 crew carried, for a total of 2,206. Of these, 493 passengers and 210 crew were saved, giving a total of 703 survivors and 1,503 non-survivors.

Most of the *Titanic* data sets list individual passengers and their main characteristics such as gender, age (or adult/child), passenger class, and survived?. Some of these give further details such as passenger name, family composition (e.g., number of siblings/spouses aboard and number of parents/children aboard), price of ticket, cabin, port of embarkment, lifeboat, body identification number, etc. Others also provide information on the crew and their relevant details, whereas some have incomplete cases removed or are split into test and training data sets. Data related to the lifeboats of the *Titanic* can be found in two of the R packages: `PASWR::titanic3` and `vcd::Lifeboats`.

## 4. Modern Graphs and Uses

After G. Bron's remarkable work, the *Titanic* data set was almost forgotten. There was basically no use for it in the following 70 years. However, in the past 40 years, many graphical methods have been used to tell the *Titanic* story as well as to illustrate some new graphical methodologies.

The use of the *Titanic* data slightly differs by discipline. In statistics, it is often used for the introduction of new graphical methods (or software) and their advantages compared to previously existing graphical methods (or software). Moreover, the data set is also used for overviews of existing graphical methods, using a well-known data set. In computer science and social sciences, the *Titanic* data set is frequently used for modeling/prediction of survival and the visualization of the results. Info Vis typically tries telling the entire story, including some data visualizations. This section gives some representative examples of the various graphical methods and uses for the *Titanic* data in statistics, computer science, and social sciences. The next section focuses on Info Vis.

### 4.1 Bar Charts

As in G. Bron's chart, the *Titanic* data is most easily displayed in bar charts. Modern uses have focused on extending this in various ways.

As one example, Hofmann (1998) used this data set to illustrate interactive methods for analyzing multivariate contingency tables. Figure 2 shows the univariate breakdowns by class, sex, age, and survive. Selecting `Survive=Yes` highlights these cases in all other graphs.

Gärtner (2017) developed *Shiny* apps for introductory statistics courses. He used the *Titanic* data as example data sets for bar charts, spine plots, mosaic plots, and common angle plots in his apps. Simple bar charts of the survival rate by class can even be found in non-academic books such as Geller (1998), p. 195.

### 4.2 Mosaic Plots and Related Graphs

G. Bron's chart illustrated some of the challenges in visualizing a table of frequencies of survival, classified by two or more factors. Mosaic plots (Hartigan and Kleiner, 1981) provide an attractive alternative, where the frequencies in a table are shown by areas of


**Figure 2**: Figure 2 from Hofmann (1998): Four barcharts with marginal distribution of the properties. Highlighted (dark) are survivors.

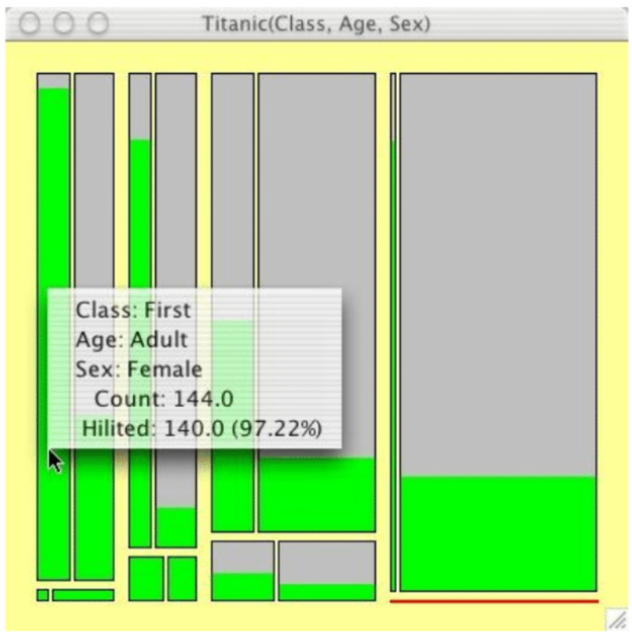



**Figure 3**: Figure 5 from Theus (2002): The Titanic Data in a Mosaic Plot.

"tiles" in a recursive partitioning of a unit square. The history of mosaic plots has been summarized in Friendly (2002).

**Basic Mosaic Plots** Theus (2002) and others illustrated the idea of an interactive mosaic plot, where the tiles for the combinations of class, age, and gender could be highlighted to show the proportion surviving in each cell. One example is shown in Figure 3.

**Visualizing Loglinear Models** Friendly (1994) described the use of mosaic plots to visualize the badness of fit of a loglinear model by shading the tiles according to the sign and magnitude of residuals in a given model. Examples of such graphs that are based on the

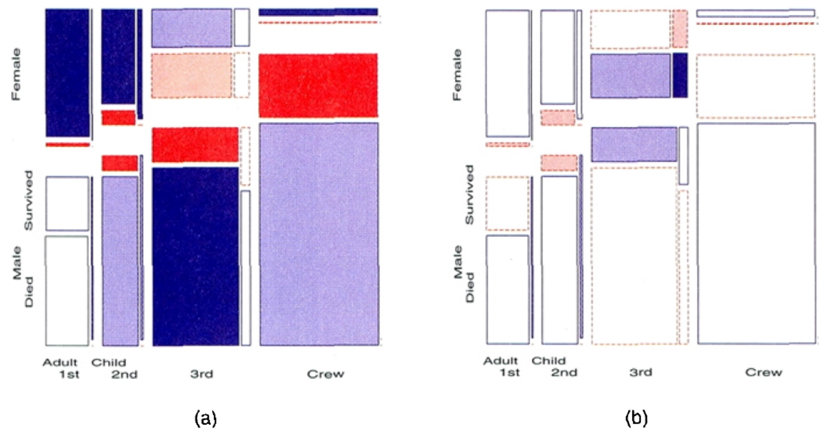



**Figure 4**: Figure 7 from Friendly (1999): Titanic data, Class, Gender, Age, and Survival: (a) joint independence; (b) main effects of Age, Gender, and Class on Survival.

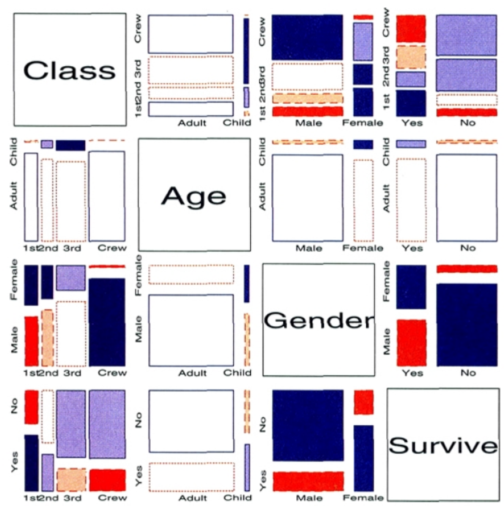



**Figure 5**: Figure 9 from Friendly (1999): Mosaic matrix of Titanic data. Each panel shows the marginal relation, fitting an independence model between the row and column variable, collapsed over other variable(s).

*Titanic* data can be found in Friendly (1999) and Friendly (2000a). One example is shown in Figure 4.

Friendly (1999) further extended the use of mosaic plots to include mosaic matrices, similar to scatterplot matrices for quantitative data. Figure 5 shows the marginal association between each pair of variables in the *Titanic* data. The row and column for `Survive` show the association of each of the predictors with survival.

**Doubledecker plots** A doubledecker plot (Hofmann, 2001) is a special case of a mosaic plot in which one variable is the response and all other are predictors. The combinations of the predictor variables are shown on the horizontal axis, with widths proportional to their joint frequencies. The response variable is split vertically. This plot displays the conditional odds ratios of the response, conditional on the predictors. It has the property that the vertical bars align when the response is unassociated with one or more predictors.

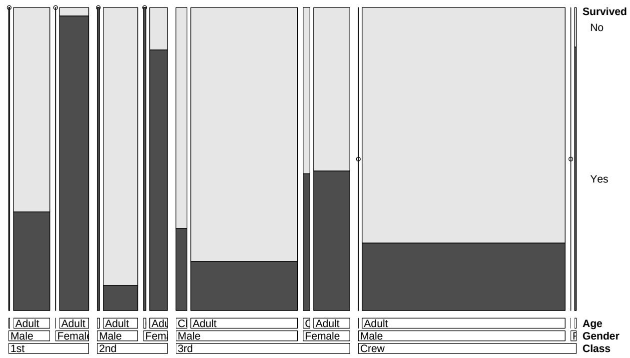



**Figure 6**: Figure 4 from Meyer et al. (2006): Double-decker plot for the Titanic data. (Figure courtesy of David Meyer, Achim Zeileis, and Kurt Hornik.)

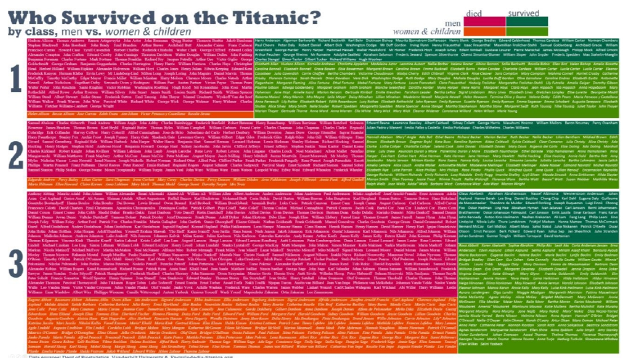



**Figure 7**: Figure 151 from Brath (2018): 1308 passengers on the Titanic, organized by class (vertically), survivorship (horizontally, serif/sans serif; red/green) and gender (plain/italic). (Figure courtesy of Richard Brath.)

Meyer, Zeileis, and Hornik (Meyer et al., 2006) described the strucplot framework for visualizing multi-way contingency tables. This provides a general framework that includes mosaic plots, association plots, doubledecker plots, and others. One of their doubledecker plots is shown in Figure 6.

**Other** Brath (2018) created mosaic plots (Figure 7) filling the tiles with the names of the victims and survivors among the 1,308 passengers. This highlights the fact that the passengers were people, not statistics. For Brath, it provided an example of how typography could be used in data visualization.

### 4.3 Parallel Sets

Parallel coordinate plots (Inselberg, 1985; Wegman, 1990) provide a way to display multi-dimensional data in 2D plots. They do this by representing the variables as a set of parallel axes, and showing each observation as a line in parallel coordinate space, rather than as a point in standard Euclidean coordinate space. Extensions of this idea for categorical data led to parallel sets plots. Jason Davies created a web page that allows to interactively explore parallel set representations of the *Titanic* data at `https://www.jasondavies.com/parallel-sets/`.

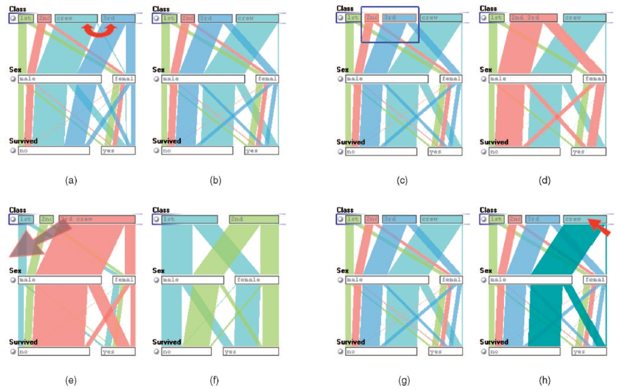



**Figure 8**: Figure 7 from Kosara et al. (2006): Basic interaction elements in Parallel Sets: reordering categories (a), (b) helps to generate a more meaningful layout; grouping categories (c), (d) enables a hierarchical analysis/exploration; excluding categories from the visualization (e), (f) allows for interactive filtering; and category highlighting (g), (h) enables the selective investigation of high-dimensional relations.

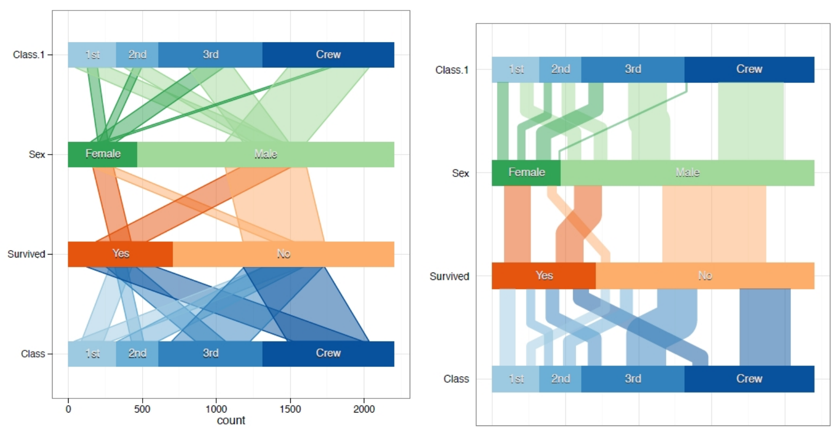
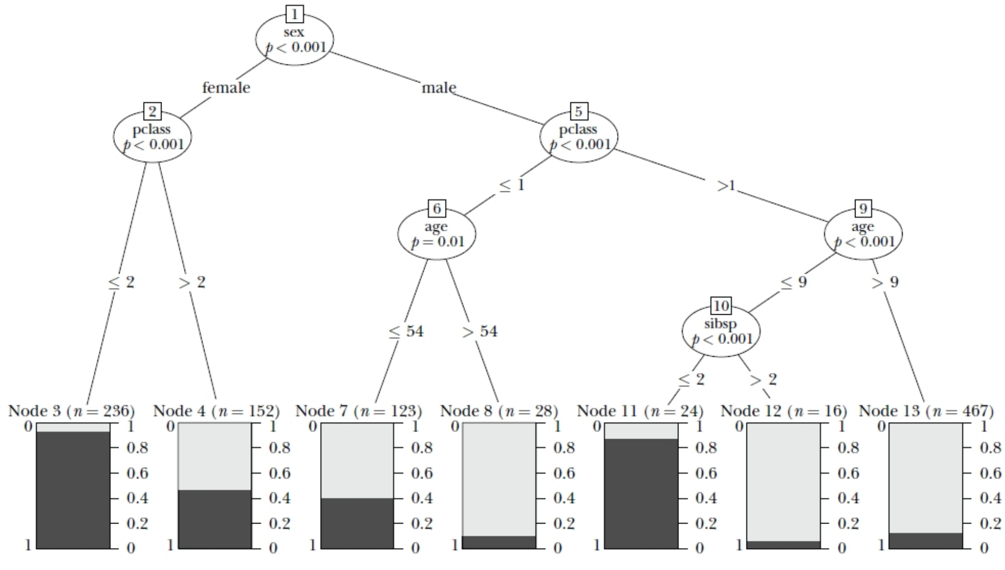



**Figure 9**: [Left] Figure 7 from Hofmann and Vendettuoli (2013): Hammock plot of the relationship between Class and Survival on the Titanic. [Right] Figure 15 from Hofmann and Vendettuoli (2013): Common angle plot of Titanic data using hammock correction.

Bendix, Kosara, and Hauser (Bendix et al., 2005) and Kosara, Bendix, and Hauser (Kosara et al., 2006) developed an interactive system to explore multivariate categorical data using parallel sets, in which the lines between categories of successive variables are of width proportional to the joint frequencies (see Figure 8).

Specific variations of parallel sets plots are called hammock plots (Schonlau, 2003) and common angle plots (Hofmann and Vendettuoli, 2013). A number of these used the *Titanic* data as examples. In particular, Hofmann and Vendettuoli (2013) pointed out that the widths of slanted lines in hammock plots (see Figure 9 [Left]) are not judged accurately. They introduced common angle plots as a perceptually-true way to better show associations of categorical variables (see Figure 9 [Right]).




*Note:* See text for interpretation.

**Figure 10**: Figure 4 from Varian (2014): A ctree for Survivors of the Titanic (black bars indicate fraction of the group that survived).

### 4.4 Tree Maps and Tree Diagrams

Cross-classified data can also be displayed as tree diagrams of various types, with branches corresponding to splits of the categories for variables in some order. Tree maps (Shneiderman, 1992) are a simple example, similar to mosaic plots. A tree map for the *Titanic* data has been posted by Robert Kosara on April 13, 2008 at `https://eagereyes.org/techniques/treemaps`.

A more powerful use of tree maps arises in connection with classification trees as models for an outcome variable such as survival. For a binary response, these are similar to a series of logistic regression models, where predictors are chosen to maximize predictive accuracy at each step. Pruning methods and cross-validation are used to control model complexity and minimize out-of-sample classification error. Varian (2014) was among the first to use the *Titanic* data set for this purpose.

Figure 10 shows the result of fitting a conditional inference tree ("ctree") predicting survival from sex, class, age and a measure of family size (`sibsp` = number of siblings plus spouse aboard). The first node splits the data by sex. The second divides by class. Among males in the right branch a third node splits by age, and those less than 9 years old are further split by `sibsp`. The bars at the bottom show the survival rate in each terminal node.

Hothorn and Zeileis (2015) introduced a more flexible framework and supporting software for fitting and visualizing tree-based models. Some of their examples are shown in Figure 11. Tree diagrams based on the *Titanic* data can also be found in the business analytics literature, e.g., in Dinsmore (2016), Figure 9-4.

### 4.5 Lifeboats Data

The *Titanic* sailed with only enough lifeboat space for half the people on board, the result of an antiquated safety code that hadn't kept pace with the growing size of ocean liners.

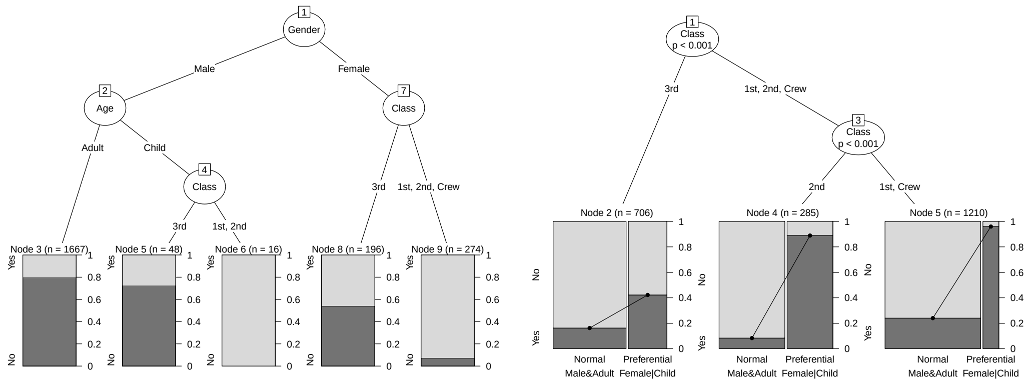



**Figure 11**: Figure 1 from Hothorn and Zeileis (2015): Tree visualizations of survival on Titanic: '**rpart**' tree converted with **as.party** and visualized by *partykit* (left); and logistic-regression-based tree fitted by **glmtree** (right). (Figure courtesy of Torsten Hothorn and Achim Zeileis.)

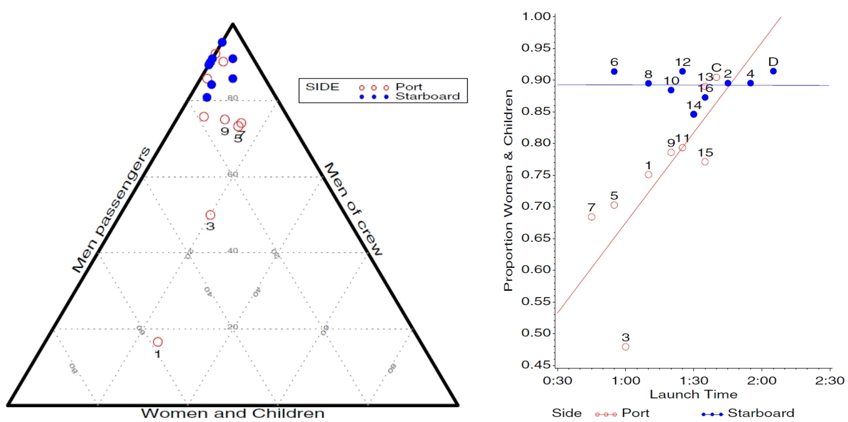



**Figure 12**: [Left] Figure 12 from Friendly (2000b): Lifeboats on the Titanic, trilinear plot. [Right] Figure 13 from Friendly (2000b): Lifeboats on the Titanic, logistic regression.

Yet some of those lifeboats were lowered less than half full. G. Bron tried to illustrate what he knew at the time, but the actual data (`vcd::Lifeboats` and `PASWR::titanic3`) allows a richer story to be told with modern graphical and statistical methods.

Friendly (2000b) displayed the lifeboat occupancy on the *Titanic* as a trilinear plot (see Figure 12 [Left]). Plotting the proportion of women and children in the lifeboats against time of launch revealed a striking difference in the regimes of loading on the port and starboard sides (see Figure 12 [Right]). Friendly and Meyer (2016) used additional graphs and statistical models to examine the loading of the lifeboats over time (see Figure 13).

### 4.6 Miscellaneous Data Graphs

The *Titanic* data also served to motivate or illustrate a wide variety of other graphical and analytic methods.

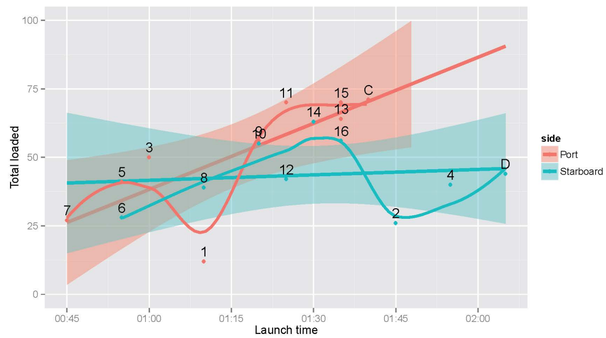



**Figure 13**: Figure 4.23 from Friendly and Meyer (2016): Number of people loaded on lifeboats on the Titanic vs. time of launch, by side of boat. The plot annotations show the linear regression and loess smooth.

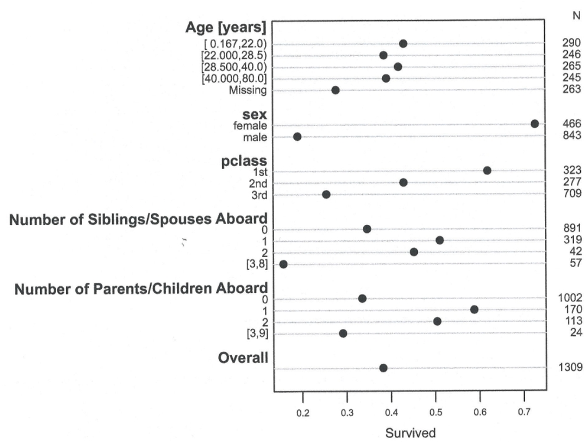



**Figure 14**: Figure 12.1 from Harrell (2015): Univariable summaries of Titanic survivors.

**Logistic Regression: Dot Plots and Nonparametric Smooths** Harrell (2015) and others used the data on the passengers in a modeling approach to predict survival from the available predictors, using logistic regression for the binary outcome (survived/died). This leads to interesting graphs showing the actual or predicted probability of survival in relation to several factors simultaneously. A basic dot plot, summarizing the probability of survival based on various predictors, is shown in Figure 14.

Nonparametric regression smooths can be use to show the relation of survival to passenger class, sex, and both. Note that "Women and children first" did not apply so well in the 3rd class as shown in Figure 15.

**Nomograms** Nomograms were originally developed by French engineer and mathematician Maurice d'Ocagne in 1891 to allow users to graphically compute the outcome of an equation without doing any calculus (d'Ocagne, 1891). Lubsen et al. (1978) extended nomograms to visualize a logistic regression model. Harrell (2001) provided implementations of logistic regression nomograms in S-Plus and briefly mentioned them in the context of the *Titanic* data (Harrell, 2001, p. 326).

Možina et al. (2004) used nomograms and interactive graphic designed to show the predicted probability of survival for various settings of the predictors in the *Titanic* data in

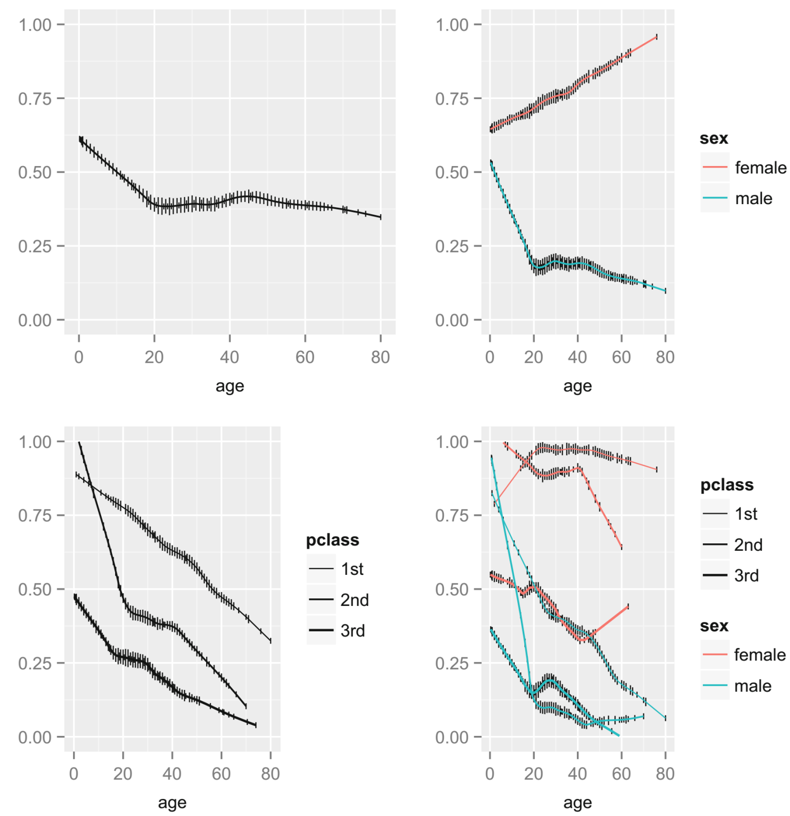



**Figure 15**: Figure 12.3 from Harrell (2015): Nonparametric regression (**loess)** estimates of the relationship between age and the probability of surviving the Titanic, with tick marks depicting the age distribution. The top left panel shows unstratified estimates of the probability of survival. Other panels show nonparametric estimates by various stratifications.

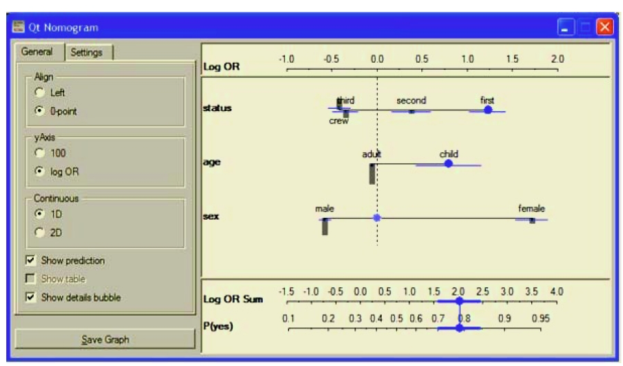



**Figure 16**: Figure 4 from Možina et al. (2004): Orange widget with the Titanic nomogram that includes confidence intervals for contributions of attribute values and class probabilities. For a woman travelling in the first class, the probability of survival is with 95% confidence between 0.87 and 0.92.

a Bayesian framework. One of their examples is shown in Figure 16.

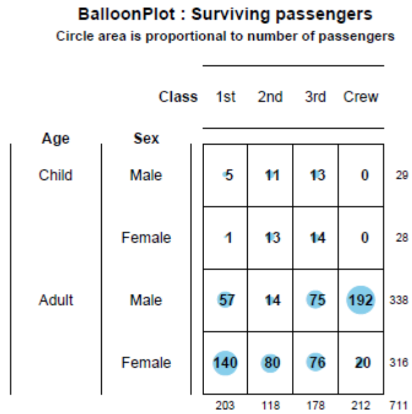



**Figure 17**: Figure 2 from Jain and Warnes (2006): Balloon plot of surviving individuals by class, gender and age. (Figure courtesy of Nitin Jain and Gregory R. Warnes.)

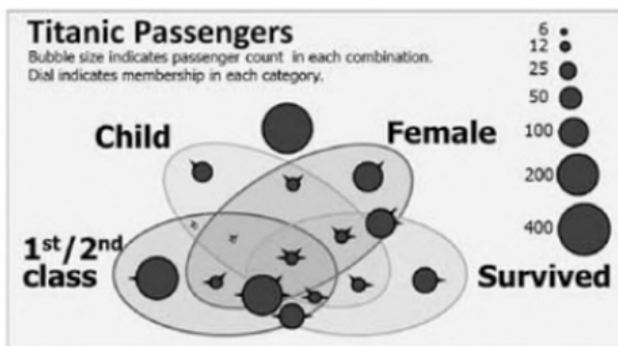



**Figure 18**: Figure 3.30 from Brath (2014): A Venn diagram of Titanic survivor data, with a bubble per segment sized to indicate the number of corresponding passengers; and with spikes per bubble to indicate set membership by pointing towards the corresponding set labels around the perimeter. For example, the large bubble near the centre bottom has three spikes, indicating that its members belong to three sets. The orientation of these spikes correspond to the location of the labels around the perimeter; therefore, based on the spikes it can be determined, this large bubble corresponds to a large number of $1^{st}/2^{nd}$ class, female passengers that survived the Titanic disaster.

**Balloon Plots** Jain and Warnes (2006) coined the term balloon plot to refer to a semi-graphic table in which the size of the cell entry was overlaid with a "balloon" to visualize the magnitudes of the values in each cell. One of their examples is shown in Figure 17.

**Venn Diagrams** Brath (2014) used an annotated Venn diagram (see Figure 18) to show the overlapping sets among combinations of the *Titanic* variables. While this diagram at first glance seems to be hard to understand, the description in the figure caption considerably helps with its interpretation.

**Steam & Leaf Plots** Brath and Banissi (2017) and Brath (2018) also extended stem & leaf plots for text visualizations. Similar to the mosaic plots created by Brath (2018) (see Section 4.2), the leaves in these stem & leaf plots are the names of the victims and

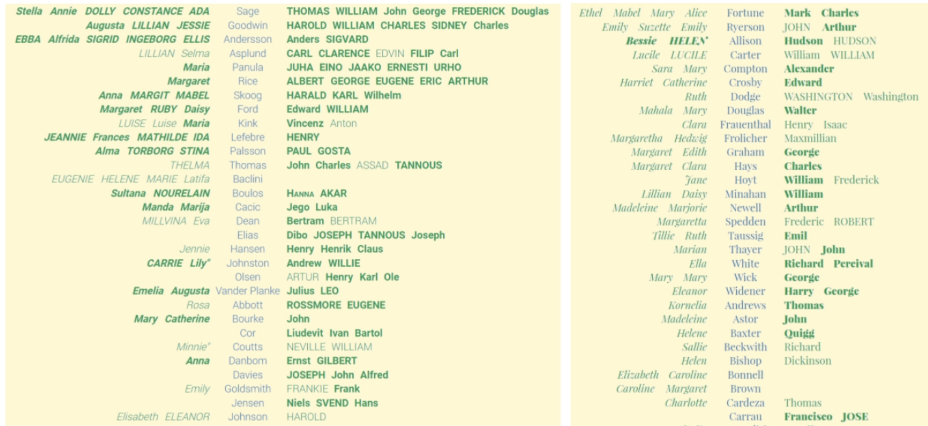


**Figure 19**: Figure 138 from Brath (2018): Titanic passengers. Left: third class families; Right: first class families. Stem indicates surname, leaf for given name, bold indicates death, italics for women, allcaps for children. (Figure courtesy of Richard Brath.)

survivors among the 1,308 passengers. Again, this highlights the fact that the passengers were people, not statistics. Two examples can be seen in Figure 19.

### 4.7 Educational Uses

Dawson (1995) used the *Titanic* data, without identifying it as such in a classroom exercise in statistical thinking: Given tables of the data on survival by class, age, gender, could students discover what the “Unusual Episode” entailed? Schumm et al. (2002) used the *Titanic* data in the classroom to illustrate the impact of social class and gender on survival.

Some textbooks made use of the *Titanic* data for specific exercises, e.g., Agresti (2007), Problem 2.7, p. 56. Others used the data for extended case studies, often covering one or more chapters. Examples include Harrell (2001), Chapter 12, Harrell (2015), Chapter 12, Theus and Urbanek (2009), Chapters 3–5 and Appendix D (with several *Titanic*-based exercises), Unwin (2015), Chapters 1, 4, 7, and 10 (with several *Titanic*-based exercises), Wilkinson (1999), Chapter 11, and Wilkinson (2005), Chapters 11 and 13.

## 5. Info Vis Applications

Just as the tragedy of the sinking of the *Titanic* inspired G. Bron to try to put the data into visual form, so too this event has been a challenge for modern graphic designers to tell the story of this disaster in ways that are both visually appealing and provide sufficient details. Unlike statistical graphs which usually focus on just one aspect, an information graphic often tries to tell the entire story all on one sheet, as in a poster presentation.

The Info Vis graphic from Barr and Johnson (2017) was a tour-de-force of visual storytelling. Shown in Figure 20 are the ones who survived (above water) and died (below water) in bar charts by age for the three passenger classes (top), the location of these classes on board the ship (center), the route (bottom third, left), the sinking and breaking apart of the *Titanic* (bottom third, center), the death rate by gender and class and also by nationality (bottom third, right), and information (time and occupancy) of the lifeboat launches (bottom).

Similarly, the Info Vis graphic from Arranz (2012) visualized the collision with the iceberg (top), sinking (right), and spread of the debris field (bottom right) of the *Titanic*. The

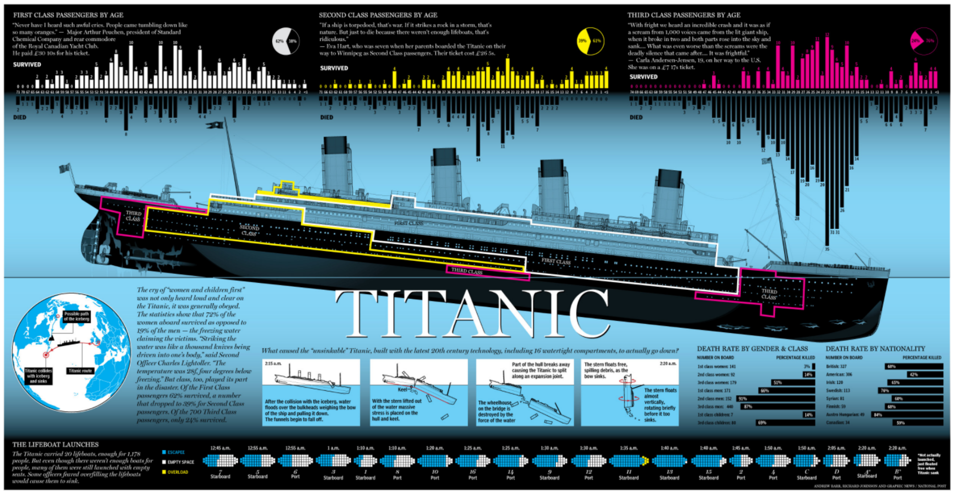



**Figure 20**: Figure from Barr and Johnson (2017).

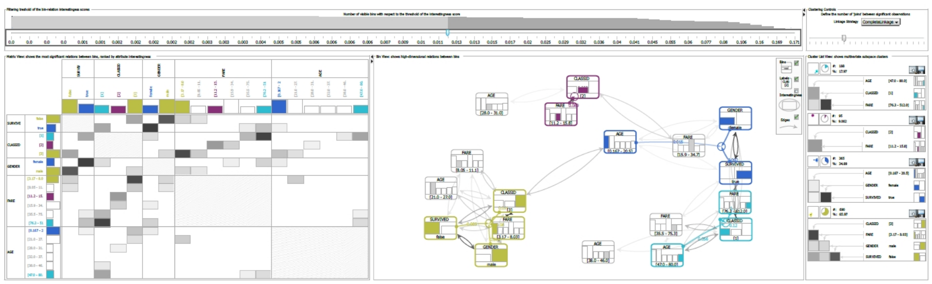


**Figure 21**: Figure 1 from Bernard et al. (2014): The Titanic data set as a proof of concept. Left: Attribute View with ranked attributes and all bin relations above the interestingness filter. Center: the Bin View aligns related bins close to each other. Right: the Cluster List View provides compact representations of clustered subspaces. The blue cluster relates to the Birkenhead Drill: 'women and children first!'.

center and bottom left contains information on water temperature, distress calls, lifeboats, the route, nearby ships and the ice field. The human toll by class, age, and gender is also shown in the center.

Bernard et al. (2014) used the *Titanic* data set as a proof of concept for the visual-interactive exploration of multivariate relations in mixed data sets. Figure 21 shows one of their examples.

Drucker and Fernandez (2015) introduced the term unit visualizations to describe a class of visualizations that explicitly represent every row in a data set. They apply this technique to produce a wide variety of unit visualizations including unit column charts, faceted small multiples, faceted scatter plots, fluctuation diagrams, summed column charts, and violin plots, all based on the *Titanic* data.

Langer and Zeiller (2017) conducted a usability study of interactive infographics in online newspapers, including *Case 1: So sank die "Titanic"*, published by Spiegel Online in 2012.

## 6. Competitions

There have been several recent competitions that used the *Titanic* data as a basis for modern statistical techniques and graphical methods. We list two highlights in this section.

**Kaggle** The *Kaggle Competition*, titled *Titanic: Machine Learning from Disaster* asked participants to predict survival on the *Titanic* and get familiar with machine learning basics. It was designed as a competition in predictive modeling, using the *Titanic* data. The data set was split into training and test samples. The goal was to devise a method to predict survival in the test sample, using only the training data set. This competition attracted nearly 10,000 teams, submitting their code, results, and commentary. Further details are accessible at `https://www.kaggle.com/c/titanic/data`. Trevor Stephens posted a tutorial for this competition on January 10, 2014, titled *Titanic: Getting Started With R* that is accessible at `https://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/`.

A few notable entries to this competition were:

- Megan Risdal: *Exploring Survival on the Titanic*, accessible at `https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic`.
- Jason: *Large Families not Good for Survival*, accessible at `https://www.kaggle.com/jasonm/large-families-not-good-for-survival`.
- Eric Bruin: *Titanic: 2nd Degree Families and Majority Voting*, accessible at `https://www.kaggle.com/erikbruin/titanic-2nd-degree-families-and-majority-voting`.

**Business Analysis Olympiad** In 2008, the city of Charlotte, North Carolina, sponsored a *Business Analysis Olympiad* to promote the business value of using visual data analysis software. It was based on the *Titanic* data and attracted teams from across the city's 14 departments to learn about new ways how to visualize and analyze this data set.

The winners were:

- First Place – "Trash Talkers" from Solid Waste Services, showing that there was significant empty space in some of the lifeboats. The analysts were Kimberly Jenkins and James Gray with Michelle Moore as their sponsor.
- Second Place – "Research Methods" from Planning, focusing on the origin and destination countries of the passengers. The analysts were Ruchi Agarwai and Evan Lowry with Steve Patterson as their sponsor.
- Third Place – "Quality CATS" from Charlotte Area Transit System, looking at the survivors by gender and class, age, and cabin. Analysts were Celia Gray and Shelly McKee with Cilia Gray as their sponsor.

Additional details can be found in the article *City of Charlotte Wows Us with Innovative "Business Analysis Olympiad"*, posted by Jock Mackinlay on November 4, 2008, at `https://www.tableau.com/blog/city-charlotte-business-analysis-olympiad`. The *Business Analysis Olympiad* and its outcome were further discussed by Stoodley (2012), p. 9.

## 7. Discussion

Overall we have located more than 40 articles and books that contain graphs based on the *Titanic* data set. Currently, there exist at least 12 R packages that host 17 different versions of this data set. Additional versions of the data can be found on the web. Numerous competitions, infographics, and single web pages made use of the *Titanic* data.

The *Titanic* nowadays can almost be considered as a huge franchise with regular new books, new movies, new TV documentaries, museums, and exhibits. Given the popularity of this overall topic and the extremely popular *Titanic* data set, one can expect to see further uses of the *Titanic* data to be continued in the future, in areas such as statistics, computer science, social sciences, and Info Vis. It may only be a question of time until further data related to the *Titanic*, such as the debris field and the description and location of objects salvaged from the ocean floor, will become publicly available in R. In fact, a comprehensive archaeological map of the site exists since 2010, as described at `https://archive.archaeology.org/1205/features/titanic_shipwreck_jean_charcot_site_map.html`.

## References


Agresti, A. (2007). *An Introduction to Categorical Data Analysis (Second Edition)*. Wiley, Hoboken, NJ.

Arranz, A. (April 15, 2012). Sinking the 'Unsinkable'. South China Morning Post. `https://multimedia.scmp.com/culture/article/SCMP-printed-graphics-memory/lonelyGraphics/201204A128.html`.

Barr, A. and Johnson, R. (April 24, 2017). Titanic. National Post. (The original appeared in 2012), `https://nationalpost.com/news/graphics/titanic-anniversary-who-was-on-the-ship-when-it-sunk-and-who-got-away`.

Bendix, F., Kosara, R., and Hauser, H. (2005). Parallel Sets: Visual Analysis of Categorical Data. In *IEEE Symposium on Information Visualization 2005, Minneapolis, MN*, pages 133–140. IEEE.

Bernard, J., Steiger, M., Widmer, S., Lücke-Tieke, H., May, T., and Kohlhammer, J. (2014). Visual–Interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum*, 33(3):291–300.

Brath, R. (2014). The Multiple Visual Attributes of Shape. In Banissi, E., Marchese, F. T., Forsell, C., and Johansson, J., editors, *Information Visualisation: Techniques, Usability and Evaluation*, pages 43–66. Cambridge Scholars Publishing, Newcastle upon Tyne.

Brath, R. (September 2018). *Text in Visualization: Extending the Visualization Design Space*. PhD thesis, London South Bank University.

Brath, R. and Banissi, E. (2017). Stem & Leaf Plots Extended for Text Visualizations. In *2017 14th International Conference on Computer Graphics, Imaging and Visualization (CGIV17), Marrakesh, Morocco*, pages 99–104. IEEE.

Bron, G. (May 4, 1912). The Loss of the "Titanic". *The Sphere*, page 103.

Dawson, R. J. M. (1995). The "Unusual Episode" Data Revisited. *Journal of Statistics Education*, 3(3). `http://jse.amstat.org/v3n3/datasets.dawson.html`.

Dinsmore, T. W. (2016). Self–Service Analytics: Hype and Reality. In Dinsmore, T. W., editor, *Disruptive Analytics: Charting Your Strategy for Next–Generation Business Analytics*, pages 199–230. Apress/Springer, New York, NY.

d'Ocagne, M. (1891). *Nomographie: Les calculs usuels effectués au moyen des abaques. Essai d'une théorie générale*. Gauthier–Villars et Fils, Paris. (In French).

Drucker, S. and Fernandez, R. (August 2015). A Unifying Framework for Animated and Interactive Unit Visualizations. Technical report. Microsoft Research, MSR–TR–2015–65, `https://www.microsoft.com/en-us/research/publication/a-unifying-framework-for-animated-and-interactive-unit-visualizations/`.

Eaton, J. P. and Haas, C. A. (1986). *Titanic, Triumph and Tragedy*. W. W. Norton & Company, New York, London.

Feldman, L. (March 12, 2018). Design: Visions of History, Told Through Art. *Time*, pages 54–55.

Friendly, M. (1994). Mosaic Displays for Multi–Way Contingency Tables. *Journal of the American Statistical Association*, 89:190–200.

Friendly, M. (1999). Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data. *Journal of Computational and Graphical Statistics*, 8(3):373–395.

Friendly, M. (2000a). *Visualizing Categorical Data*. SAS Publishing, Cary, NC.

Friendly, M. (2000b). Visualizing Categorical Data: Data, Stories, and Pictures. In *Proceedings of the Twenty–Fifth Annual SAS Users Group International Conference (SUGI 25), Indianapolis, IN*. SAS Institute Inc.

Friendly, M. (2002). A Brief History of the Mosaic Display. *Journal of Computational and Graphical Statistics*, 11(1):89–107.

Friendly, M. and Meyer, D. (2016). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. CRC Press/Taylor & Francis, Boca Raton, FL.

Friendly, M., Symanzik, J., and Onder, O. (2019). Visualising the Titanic Disaster. *Significance*, 16(1):14–19.

Gärtner, J. (2017). Programming and Evaluation of Shiny Applications for Lectures. Master's thesis, Humboldt-Universität zu Berlin, School of Business and Economics.

Geller, J. B. (1998). *Titanic: Women and Children First*. W. W. Norton & Company, New York, London.

Harrell, Jr., F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, NY.

Harrell, Jr., F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (Second Edition)*. Springer International Publishing, Switzerland.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for Contingency Tables. In Eddy, W. F., editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273. Springer–Verlag, New York, NY.

Hofmann, H. (1998). Simpson on Board the Titanic? Interactive Methods for Dealing with Multivariate Categorical Data. *Statistical Computing and Statistical Graphics Newsletter*, 9(2):16–19.

Hofmann, H. (2001). Generalized Odds Ratios for Visual Modeling. *Journal of Computational and Graphical Statistics*, 10(4):628–640.

Hofmann, H. and Vendettuoli, M. (2013). Common Angle Plots as Perception–True Visualizations of Categorical Associations. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2297–2305.

Hothorn, T. and Zeileis, A. (2015). *partykit*: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research*, 16(118):3905–3909.

Inselberg, A. (1985). The Plane with Parallel Coordinates. *The Visual Computer*, 1:69–91.

Jain, N. and Warnes, G. R. (2006). Balloon Plot: Graphical Tool for Displaying Tabular Data. *R News, The Newsletter of the R Project*, 6(2):35–38. `http://CRAN.R-project.org/doc/Rnews/Rnews_2006-2.pdf`.

Kosara, R., Bendix, F., and Hauser, H. (2006). Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568.

Langer, J. and Zeiller, M. (2017). Evaluation of the User Experience of Interactive Infographics in Online Newspapers. In Aigner, W., Moser, T., Blumenstein, K., Zeppelzauer, M., Iber, M., and Schmiedl, G., editors, *FMT 2017, Proceedings of the 10th Forum Media Technology and 3rd All Around Audio Symposium, St. Pölten, Austria*, pages 97–106. Forum Media Technology 2017, Vol–2009. `http://ceur-ws.org/Vol-2009/`.

Lubsen, J., Pool, J., and van der Does, E. (1978). A Practical Device for the Application of a Diagnostic or Prognostic Function. *Methods of Information in Medicine*, 17(2):127–129.

Meyer, D., Zeileis, A., and Hornik, K. (2006). The Strucplot Framework: Visualizing Multi–Way Contingency Tables with vcd. *Journal of Statistical Software*, 17(3):1–48. `http://www.jstatsoft.org/v17/i03/`.

Možina, M., Demšar, J., Kattan, M., and Zupan, B. (2004). Nomograms for Visualization of Naive Bayesian Classifier. In Boulicaut, J.-F., Esposito, F., Giannotti, F., and Pedreschi, D., editors, *Knowledge Discovery in Databases: PKDD 2004, 8th European Conference on Principles and*

*Practice of Knowledge Discovery in Databases, Pisa, Italy*, pages 337–348, Berlin, Heidelberg. Springer.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rendgen, S. and Wiedemann, J. (Ed.) (2012). *Information Graphics*. Taschen GmbH, Köln.

Schonlau, M. (2003). Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots. In *2003 JSM Proceedings*, Alexandria, VA. American Statistical Association. (CD).

Schumm, W. R., Webb, F. J., Castelo, C. S., Akagi, C. G., Jensen, E. J., Ditto, R. M., Spencer-Carver, E., and Brown, B. F. (2002). Enhancing Learning in Statistics Classes through the Use of Concrete Historical Examples: The Space Shuttle Challenger, Pearl Harbor, and the RMS Titanic. *Teaching Sociology*, 30(3):361–375.

Shneiderman, B. (1992). Tree Visualization with Tree–Maps: A 2–D Space–Filling Approach. *ACM Transactions on Graphics*, 11(1):92–99.

Stoodley, N. (2012). Democratic Analytics: A Campaign to Bring Business Intelligence to the People. *Business Intelligence Journal*, 17(1):7–12.

Symanzik, J., Friendly, M., and Onder, O. (August 30, 2018). 100+ Years of Graphs of the Titanic Data. CompStat 2018, Iasi, Romania. http://www.math.usu.edu/~symanzik/talks/2018_ComStat.pdf.

Symanzik, J., Friendly, M., and Onder, O. (July 30, 2019). The Unsinkable Titanic Data. Joint Statistical Meetings, Denver, Colorado. http://www.math.usu.edu/~symanzik/talks/2019_asa.pdf.

Theus, M. (2002). Interactive Data Visualization Using Mondrian. *Journal of Statistical Software*, 7(11). http://www.jstatsoft.org/v07/i11/.

Theus, M. and Urbanek, S. (2009). *Interactive Graphics for Data Analysis: Principles and Examples*. Chapman & Hall/CRC, Boca Raton, FL.

Unwin, A. (2015). *Graphical Data Analysis with R*. CRC Press/Taylor & Francis, Boca Raton, FL.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wegman, E. J. (1990). Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, 85:664–675.

Wilkinson, L. (1999). *The Grammar of Graphics*. Springer, New York, NY.

Wilkinson, L. (2005). *The Grammar of Graphics (Second Edition)*. Springer, New York, NY.