In Search of the Royal Mail Ship (RMS) Titanic

Chuck Cartledge

April 23, 2020

Contents

List of Tables

List of Figures

1 Introduction

The sinking of the Royal Mail Ship (RMS) Titanic on her maiden voyage is a source of constant mystery and romance. Now after more than a century, there are still unanswered questions about the disaster that made her a part of the English lexicon. Perhaps the simplest question is: how many people (passengers and crew) were on board when she sank, and how many survived? Surprisingly, there is no definitive answer to these, the most simple of questions. Neither from the White Star Line (her owner), nor from the British Wreck Commissioner assigned to inquiry into her sinking. In this report, we will enumerate some of the disparate sources, and look at some data that has made its way into the R programming language.

2 Discussion

The RMS Titanic (see Figure 1 on the following page), set sail from South Hampton England for New York, New York on 10 April 1912. She had a fire in her coal stores that started almost 10 days before she sailed, and continued after leaving South Hampton[12]. She had 16 lifeboats, and four Engelhardt "collapsibles" (see Figure 2 on page 3) that could collectively hold 1,178 people (aka, souls)[8]. When the Titanic set sail, she had approximately 2,224 passengers and crew. Titanic could had a maximum capacity of 3,327 souls. The number of survivors differs between sources; ranging from 706[10, 11] to 712[7]. 711 are reported to have survived according to the R library `titanic`.

Titanic's sailing was such a special event, that special booklets were created (see Figure 3 on page 4), complete with list of first class passengers[2].

2.1 Sources of data

One of the many things that is unusual about the sinking of the Titanic, is that there does not seem to be a definitive list of passengers and crew, let alone a list that identifies those who survived and those who did not[2]. In the following sections, we will investigate a few of these sources.

2.1.1 R

A summary of the classes of people on the Titanic, and whether or not they survived, is part of the standard R installation in the table `Titanic` (see Figure 4 on page 5).

> *"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."*
>
> R Staff [9]

[2] `http://www.phillyseaport.org/web_exhibits/mini_exhibits/titanic_passenger_list/titanic_passenger_list-object-passenger_list.html`

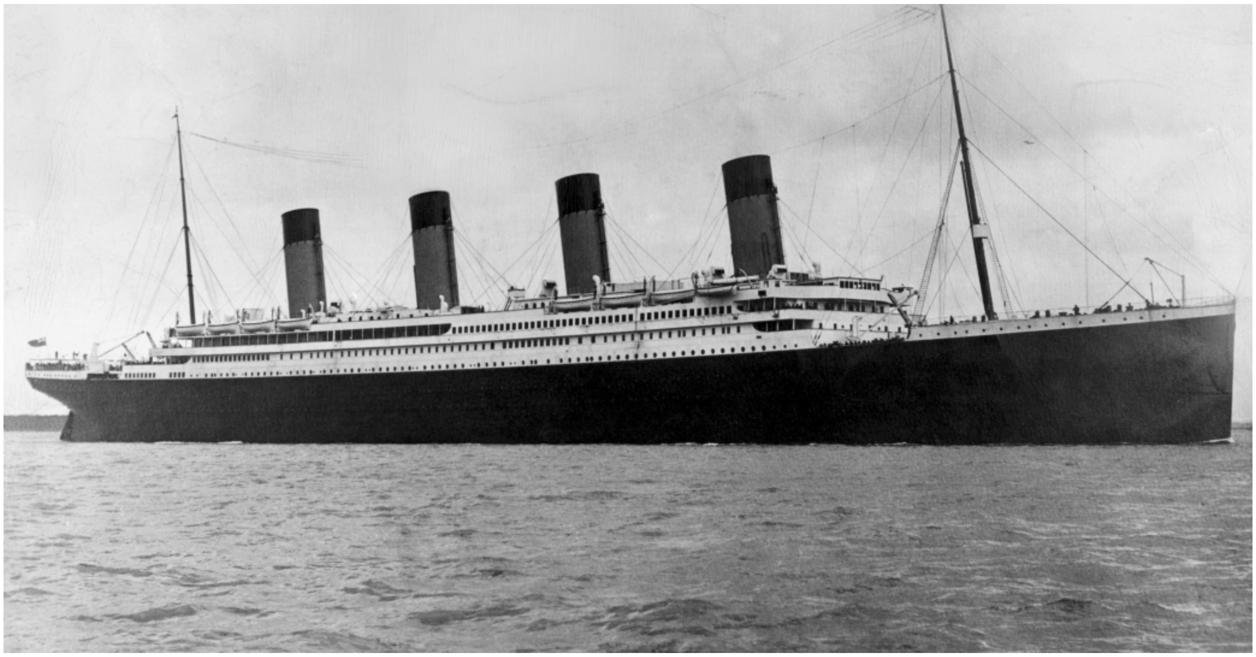

Figure 1: The RMS Titanic. Photographed 10 April 1912[1]. The Titanic sank 15 April 1912.

2.1.2 R package "titanic"

A collection of 1,309 passenger records organized to support the Kaggle competition[3]. The total data is divided into training and testing portions to support machine learning. The data does not have any crew data.

> *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. Whereas the base R Titanic data found by calling data("Titanic") is an array resulting from cross-tabulating 2201 observations, these data sets are the individual non-aggregated observations and formatted in a machine learning context with a training sample, a testing sample, and two additional data sets that can be used for deeper machine learning analysis.*
>
> Hendricks [4]

The library has copies of the training and testing dataset used by the Kaggle competition to validate and test various machine learning algorithms. While the `titanic_train` data set has which passenger survived or not, the `titanic_test` data set does not. So the only practical way to see how well your approach worked was to submit your R script to Kaggle and the await results. There is sufficient information in the `titanic_test` data set to reconstruct who lived or died, it may not be worth the effort.

2.1.3 R package "vcdExtra"

A collection of 1,309 passenger records.

> *Provides additional data sets, methods and documentation to complement the 'vcd' package for Visualizing Categorical Data and the 'gnm' package for Generalized Nonlinear Models.*

[3] https://www.kaggle.com/c/titanic

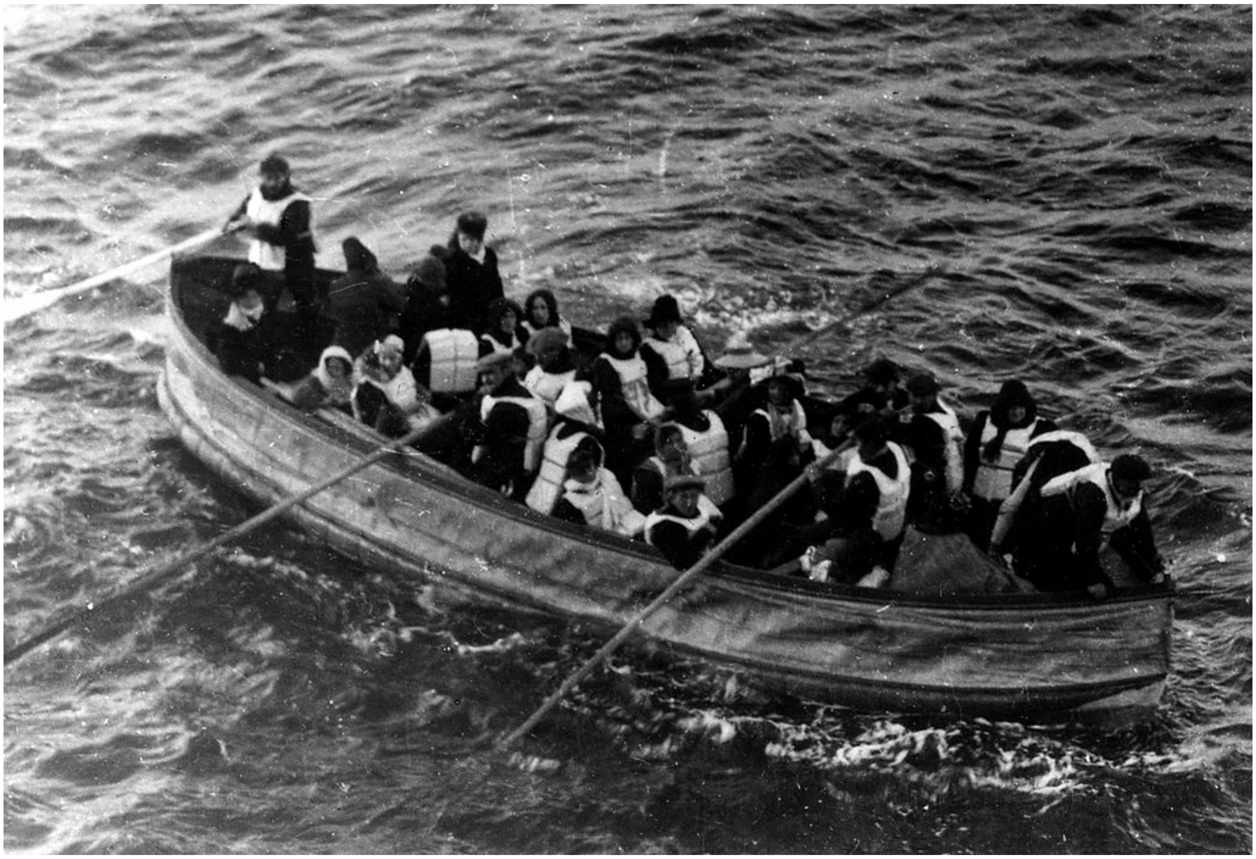

Figure 2: Titanic's Collapsible Boat D approaches RMS Carpathia.

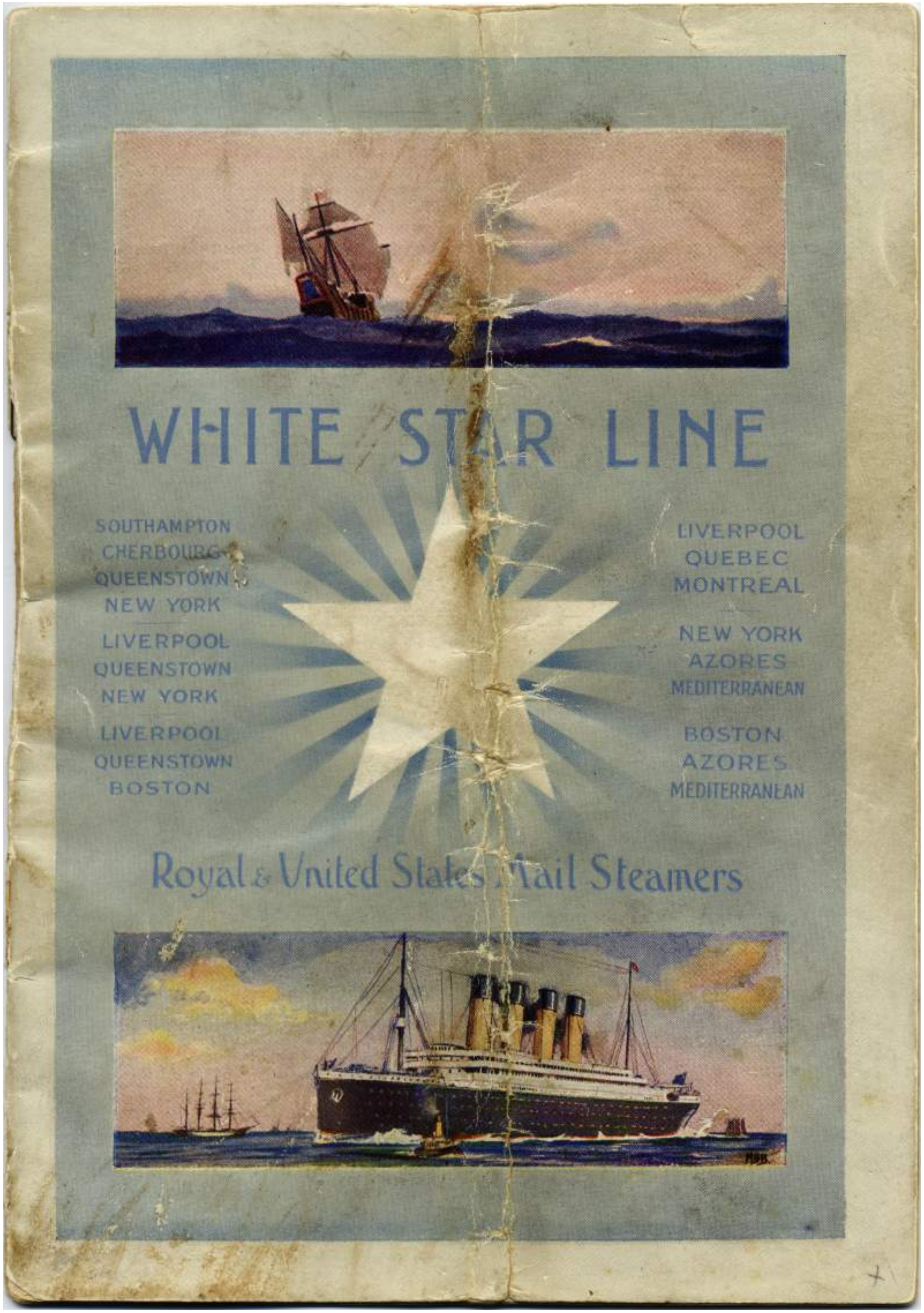

Figure 3: Commemorative sailing booklet.

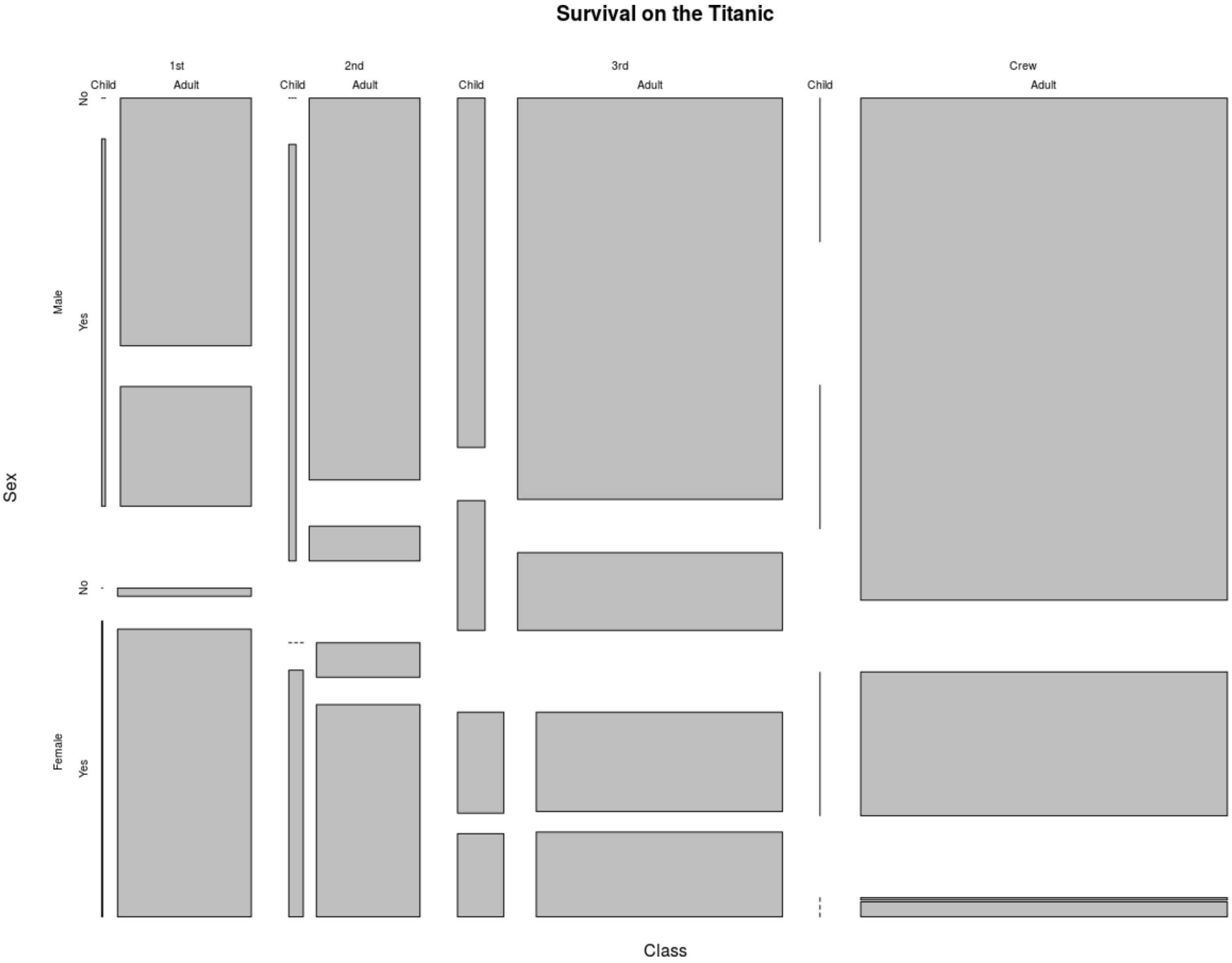

Figure 4: A Mosaic plot of R Titanic data. A summary of all personnel on the RMS Titanic broken down by gender, by survival or not, and class. It is interesting to look at the data and consider the adage: “women and children first.”

Friendly et al. [3]

The vcdExtra::Titanicp is a data frame with 1,309 observations on the following 6 variables (see Figure 5 on the following page):

- *class* a factor with levels 1^{st}, 2^{nd}, and 3^{rd}
- *survived* a factor with levels died, and survived (1 and 2 respectively)
- *sex* a factor with levels female, and male
- *age* passenger age in years (or fractions of a year, for children), a numeric vector; age is missing for 263 of the passengers
- *sibsp* number of siblings or spouses aboard, integer: 0:8
- *parch* number of parents or children aboard, integer: 0:6

In many ways, "vcdExtra::Titanicp" is a union of the "titanic::titanic_train" and "titanic::titanic_test" datasets, less some of the columnular values. The vcdExtra::Titanicp data supports looking at the data in different ways[3] (see Figure 6 on page 8). If we use all the data columns, then the decision tree becomes much more interesting

2.1.4 The crew

The previous sections focused on Titanic's passengers, but the ship also had crew and almost all Titanic lists ignore them. I was able to find one site that claimed to have a list of crew members, their job, and whether they survived or not. Port Cities Southampton[4] claims to be a digital archive of maritime activities for the Port of Southampton, including the departure of the Titanic[5]. The crew list is broken into several parts based on last names, and is available for download as PDF files.

An R script was written to parse the crew list files and put the data into a data frame compatible with the other Titanic datasets.

2.2 Questions that can be asked

Every person that sailed on the Titanic had a long list of attributes that could be used to describe them. These attributes include:

- Class (1^{st}, 2^{nd}, 3^{rd}, crew, other)
- Age,
- Fare,
- Gender,
- Number of traveling family members,
- Number of siblings,
- Place of embarkation, or

[4] http://www.plimsoll.org/StartHere/AboutUs/default.asp

[5] http://www.plimsoll.org/Southampton/Titanic/titaniccrewlist/Default.asp

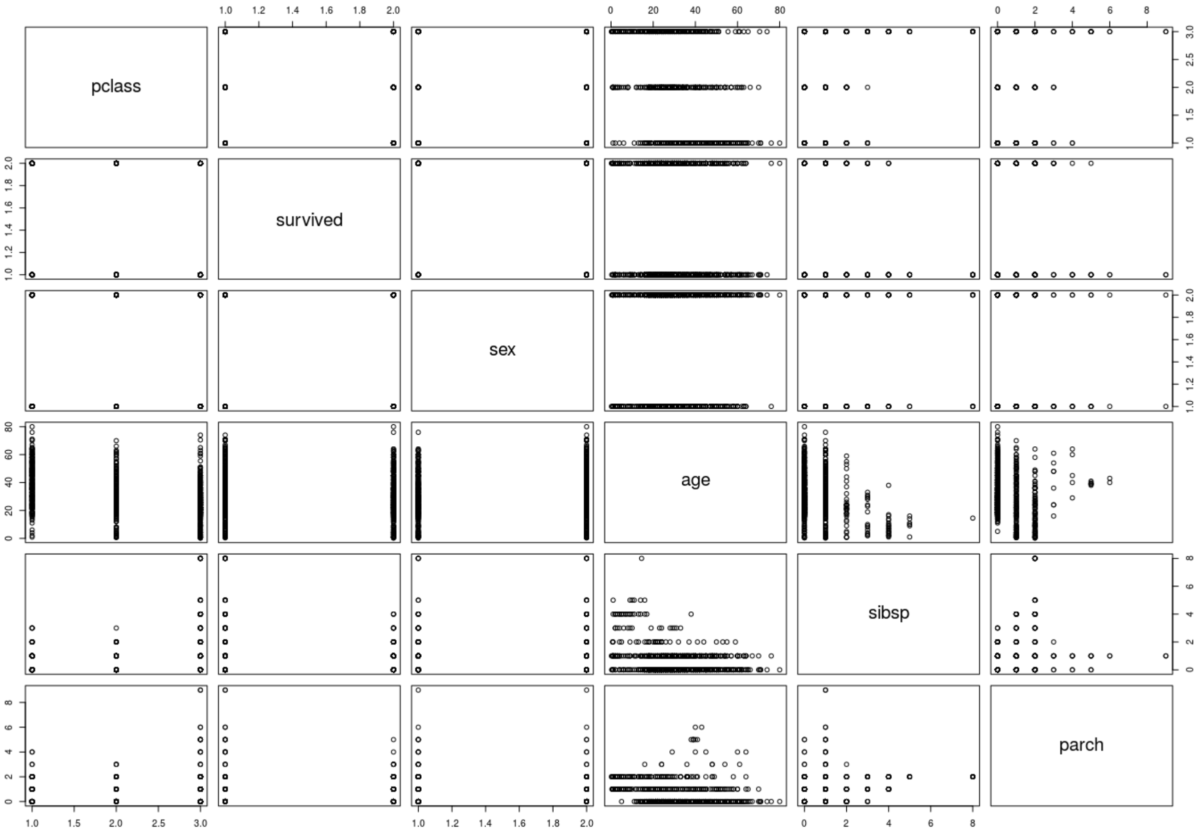

Figure 5: vcdExtra Titanic data. The data shows that only three classes of data are present (no crew members), that everyone survived or not, that there were only 2 genders identified, and that the attributes of age, sibsp, and parch have the most variability.

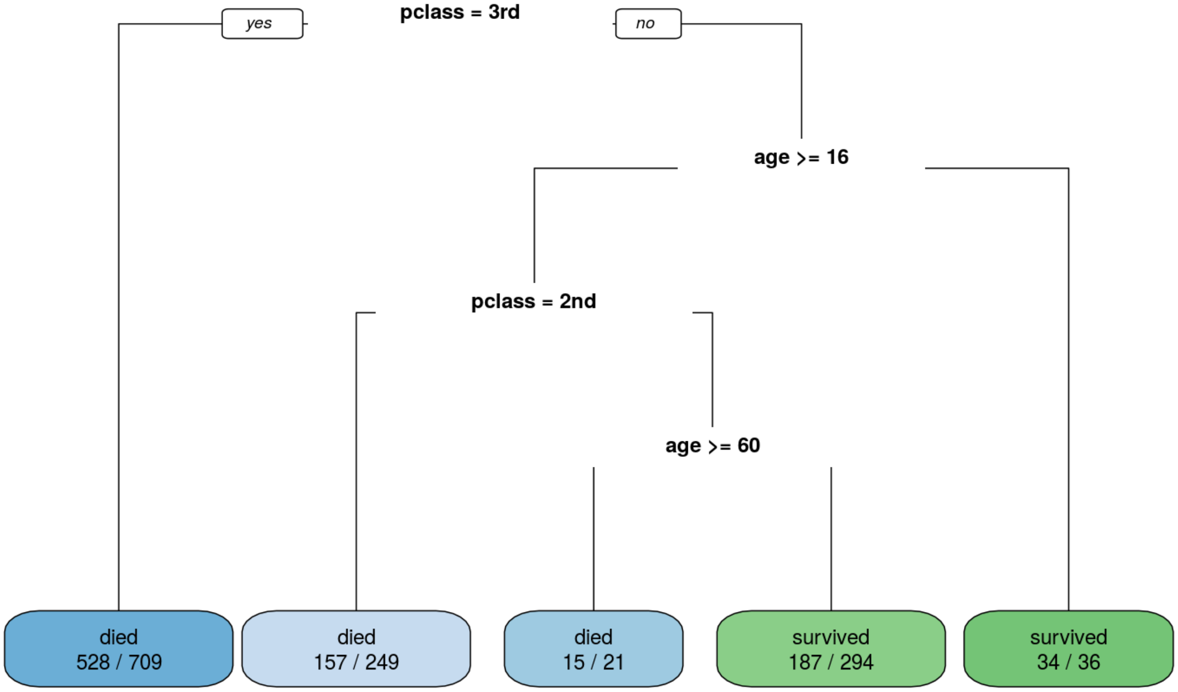

Figure 6: vcdExtra Titanic data decision tree based on passenger class, age, and number and type of traveling companions. Based on the data; if you were not in 3^{rd} class, were, under 16, and in 2^{nd} class, then 157 of the 249 people like you died.

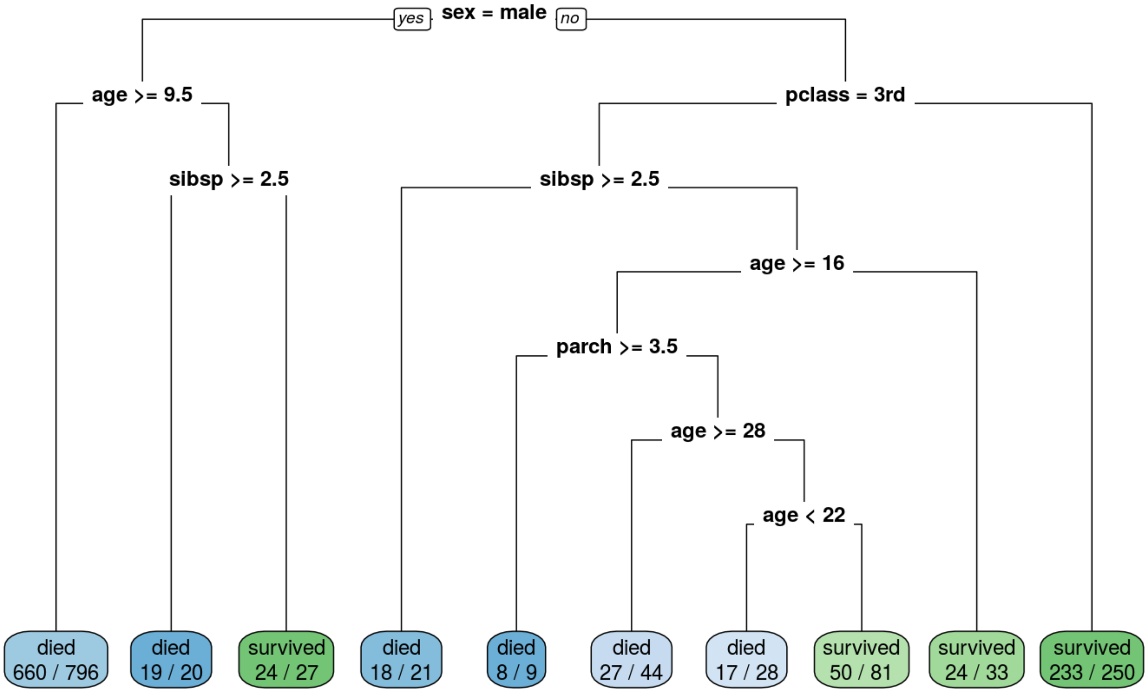

Figure 7: vcdExtra Titanic data decision tree based on all available data.

Table 1: Command line, or pass parameters. The `predictions.R` R script supports a variety of pass parameters, or command line arguments.

Name	Switch	Default	Meaning
sourceSelection	s	internal	Which dataset to use.
trainingPercentage	t	30	Percentage of data (number of people) to use in the training set. The remainder will be used as the test set.
dataCollection	d	FALSE	Should a long run be made to collect data based on the data set selected.
verbose	v	FALSE	Control lots of debugging information.

- Nationality of each person (there have been comments that percentage wise Americans survived than non-Americans because the Americans believed the announcements to abandon ship),

Not all of the datasets included in `predictions.R` have all attributes. The internal dataset is missing the attribute “embarkation” (see Figure 8 on the following page). While the optional dataset, has all values (see Figure 9 on page 12).

Each of these attributes can be used to form the question as to who survived and who did not. Not all datasets have all attributes.

For our investigation, the question is:

What is the likelihood that someone survived based on their gender, the number of siblings on board, the number of people in the traveling family unit, and the group’s traveling class (1st, 2nd, 3rd, and crew).

2.3 Presentation of results

An R script was written to compare the different datasets and to use different partitioning approaches on the datasets. The script supports a collection of arguments either via the command line, or by passing arguments via the `main()` function (see Table 1).

`predictions.R` incorporates four different partitioning algorithms (one algorithm has two variants). They are:

rpart from the **rpart** library. “Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone.”[13]

C50 from the **C50** library. “C5.0 decision trees and rule-based models for pattern recognition.” [6]

Random Forest from the **randomForest** library. “Classification and regression based on a forest of trees using random inputs.”[1]

J48 from the **RWeka** library. “An R interface to Weka (Version 3.9.1). Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization.”[5] The J48 algorithm is run in a pruned and unpruned mode.

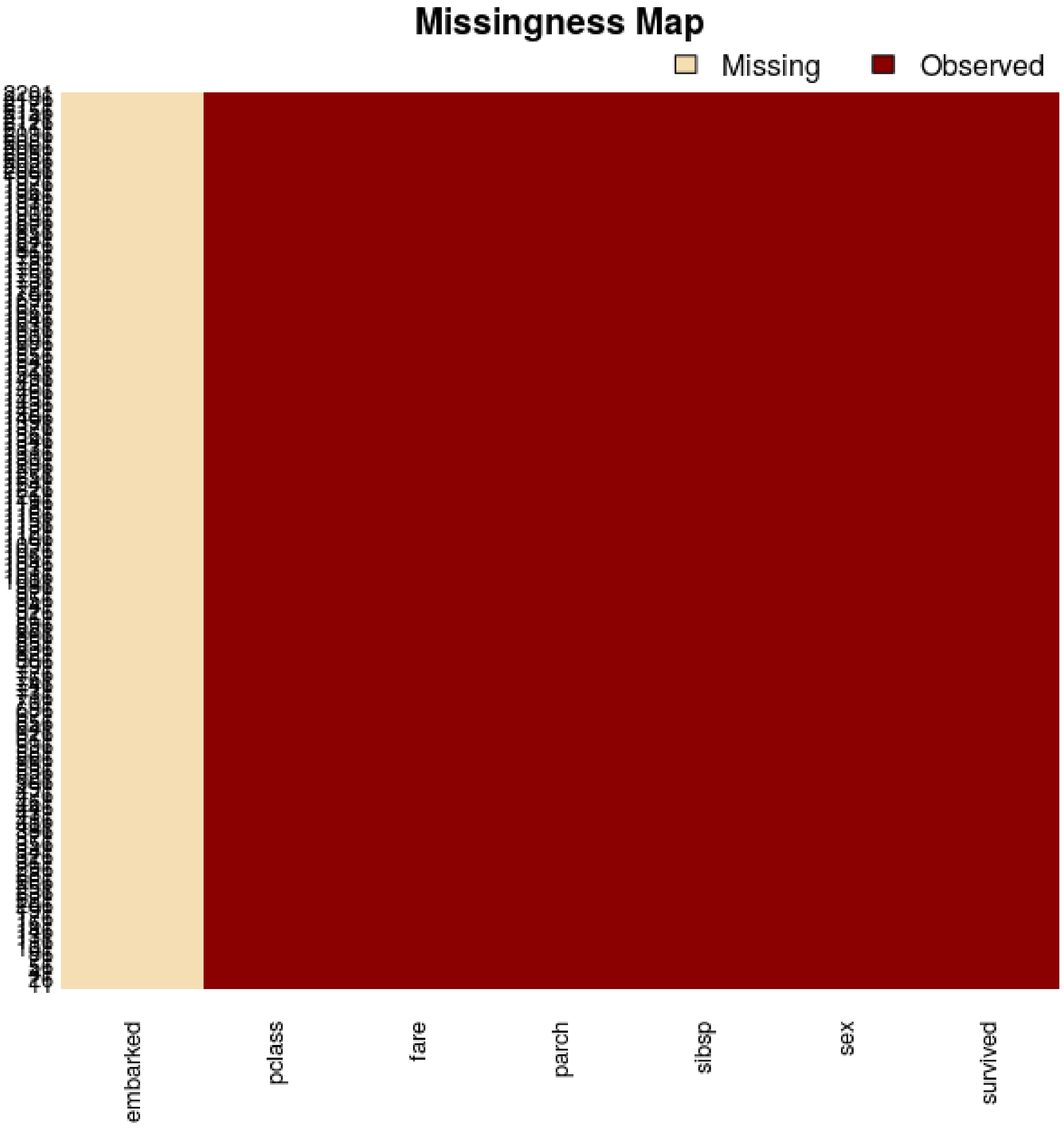

Figure 8: Missing data map for the “internal” dataset.

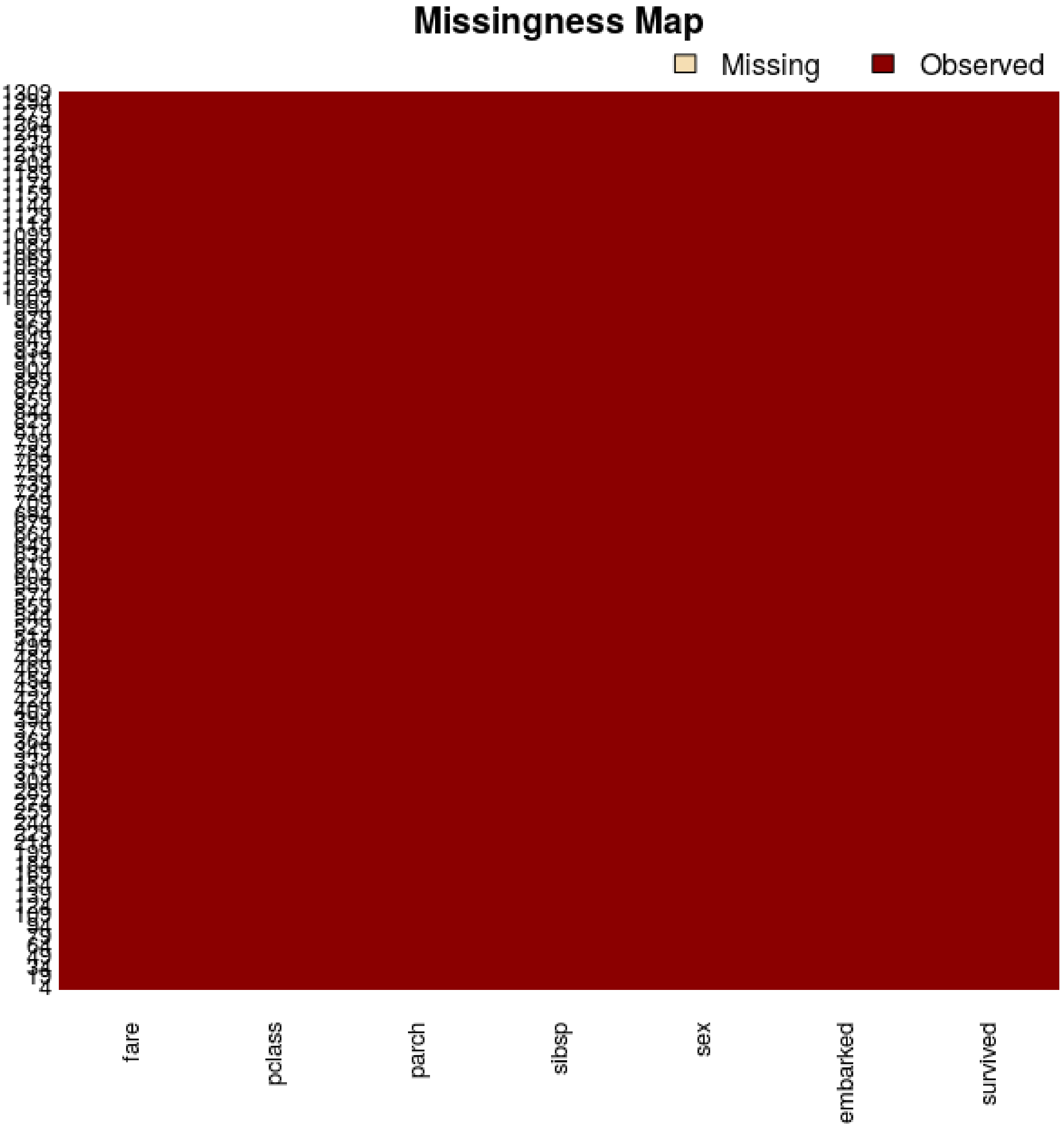

Figure 9: Missing data map for the “titanic3” dataset.

Table 2: Cross comparison of different partitioning algorithms. The row names correspond to the cells in a confusion matrix, with TRUE positives and negatives, and FALSE positives and negatives.

	C50	**Random Forest**	**rpart**	**J48 (unpruned)**	**J48 (pruned)**
True +	1031	1031	1031	1031	1031
True -	175	175	175	175	175
False +	12	12	12	12	12
False -	323	323	323	323	323
Accuracy	0.7826087	0.7826087	0.7826087	0.7826087	0.7826087
Kappa	0.4061712	0.4061712	0.4061712	0.4061712	0.4061712
	(see Figure 10 on page 15)	(see Figure 11 on page 16)	(see Figure 12 on page 17)	(see Figure 13 on page 17)	(see Figure 14 on page 18)

Each algorithm was run using the default program values and **verbose = TRUE** so that cross algorithm comparisons can be made (see Table 2). Data on each algorithm is reported:

True + from a confusion matrix, this is the TRUE Positive.

True - from a confusion matrix, this is the TRUE Negative.

False + from a confusion matrix, this is the FALSE Positive.

False - from a confusion matrix, this is the FALSE Negative.

Accuracy the accuracy based on data in the confusion matrix $Accuracy = \frac{truePositive+trueNegative}{allPositive+allNegative}$

Kappa

While the figures created by the default values are interesting (see Figures 10 on page 15 through 14 on page 18), they only provide a snapshot of how their respective algorithms perform based on a single partitioning of the data into training and testing subsets. It might be more informative to see how the accuracy of the algorithms vary as a function of the size of the training dataset. To answer that question, `predictions.R` was run with `dataCollection = TRUE` and `trainingPercentage = 1`. This causes the script to vary `trainingPercentage` from 1 to 90%, in 10% steps. The accuracy for each algorithm was captured and plotted (see Figure 15 on page 19). Based on the collected data, it appears that training percentages from about 10 to 60 result in all algorithms having nearly identical accuracies. Below 10%, the Random Forest approach appears best. Above 70%, they all seem to be an overfit the data.

3 Conclusion

A number of interesting things that came out of this investigation, including:

1. The Titanic carried more lifeboats than law required, but far too few to save all personnel (1,178 versus 3,327).
2. There does not appear to be an official and agreed to number of people who sailed on the RMS Titanic, and of those who survived or died.
3. The values from in built in R data set (`datasets:Titanic`)[6] is a reasonable approximation, although it has limited attributes.

[6] At the R command prompt: `library(help=datasets)`

4. There are many Titanic people data sets, but some do not contain all classes of people (1^{st}, 2^{nd}, 3^{rd}, and crew).
5. Various passenger and crew lists (with more attributes than the built in Titanic dataset) are available, and can be consolidated into a dataset with more attributes, and still have approximately the same number of people.
6. All of the decision tree algorithms tested had comparable results (˜76% accuracy) when the training dataset was between 10 and 60% of the entire dataset.
7. Random forest performed most consistently over the widest range of training percentages of all tested algorithms.

In summary: most of the Titanic personnel datasets are very comparable, and the random forest decision tree consistently worked the best.

References

[1] Leo Breiman, *randomforest: Breiman and cutlers random forests for classification and regression*, http://stat-www.berkeley.edu/users/breiman/RandomForests, 2006.

[2] RJM Dawson, *The "Unusual Episode" Data Revisited*, Journal of Statistics Education **3** (1995), no. 3, 1–7.

[3] Michael Friendly, Heather Turner, Achim Zeileis, and Maintainer Michael Friendly, *Package 'vcdExtra'*, (2016).

[4] Paul Hendricks, *Titanic Passenger Survival Data Set*, https://github.com/paulhendricks/titanic, 2015.

[5] K. Hornik, A. Zeileis, T. Hothorn, and C. Buchta, *RWeka: an R interface to Weka*, R package version 0.4-32 (2017).

[6] M. Kuhn, S. Weston, N. Coulter, M. Culp, and R. Quinlan, *C5.0 Decision Trees and Rule-Based Models*, R Package Version 0.1. 0 **24** (2015).

[7] Encyclopedia Titanica Staff, *Titanic Survivors*, https://www.encyclopedia-titanica.org/titanic-survivors/, 2017.

[8] History Staff, *Titanic*, http://www.history.com/topics/titanic, 2017.

[9] R Staff, *Survival of passengers on the Titanic*, package:datasets, 2017.

[10] Titanic Facts Staff, *Titanic Survivors*, http://www.titanicfacts.net/titanic-survivors.html, 2017.

[11] Titanic Universe Staff, *How Many People Survived the Titanic Disaster?*, http://www.titanicuniverse.com/how-many-people-survived-the-titanic-disaster/1253, 2017.

[12] Wikipedia Staff, *British Wreck Commissioner's inquiry into the sinking of the RMS Titanic*, https://en.wikipedia.org/wiki/British_Wreck_Commissioner%27s_inquiry_into_the_sinking_of_the_RMS_Titanic, 2017.

[13] Terry Therneau, Beth Atkinson, and Brian Ripley, *rpart*, Available at CRAN. R-project. org/package= rpart. Accessed May (2015).

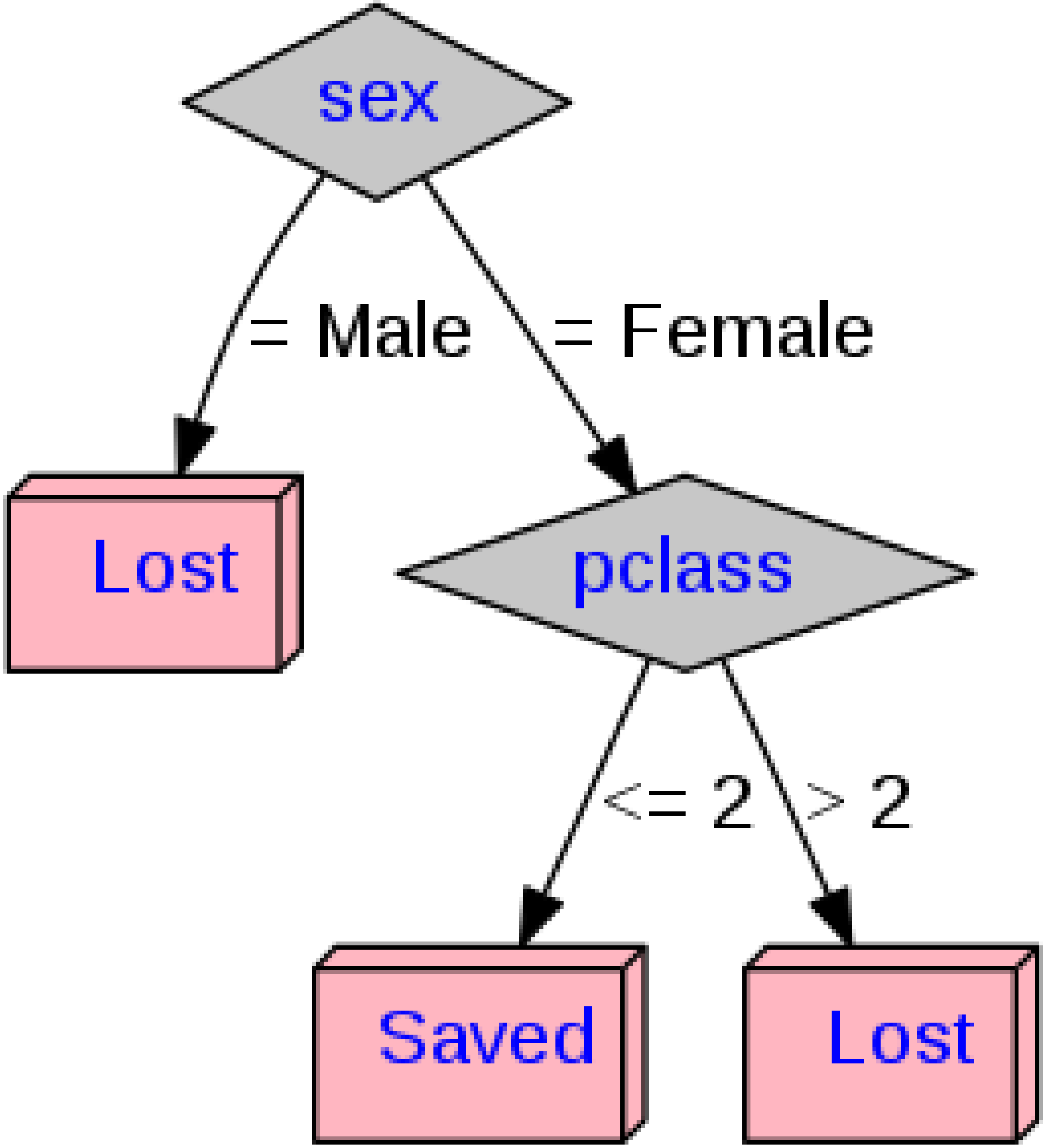

Figure 10: C50 decision tree.

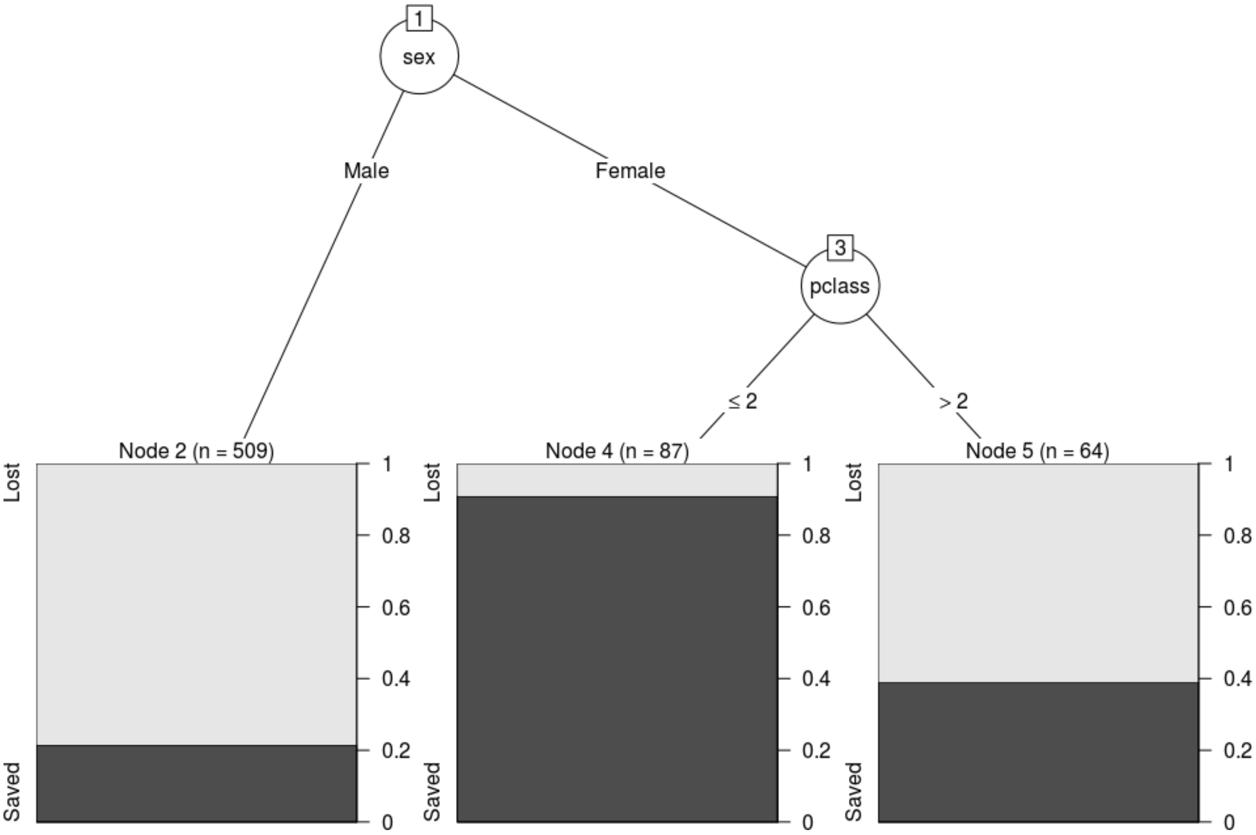

Figure 11: Random Forest decision tree.

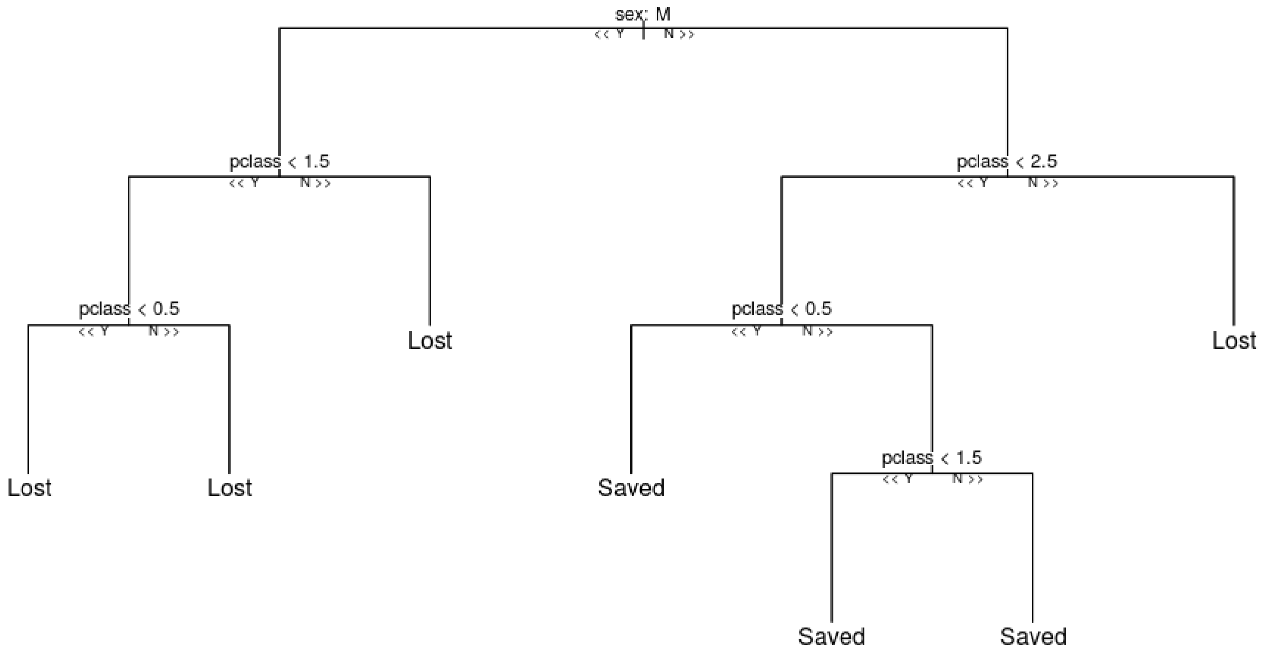

Figure 12: rpart decision tree.

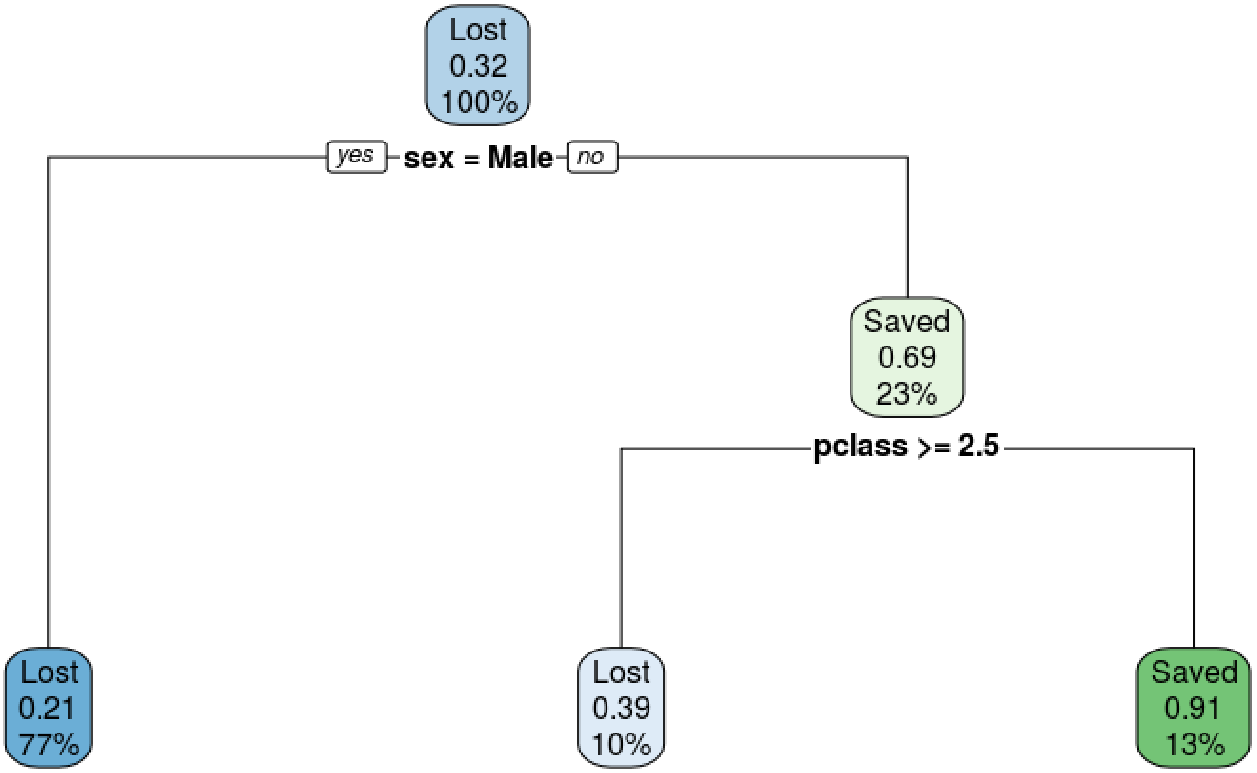

Figure 13: J48 (unpruned) decision tree.

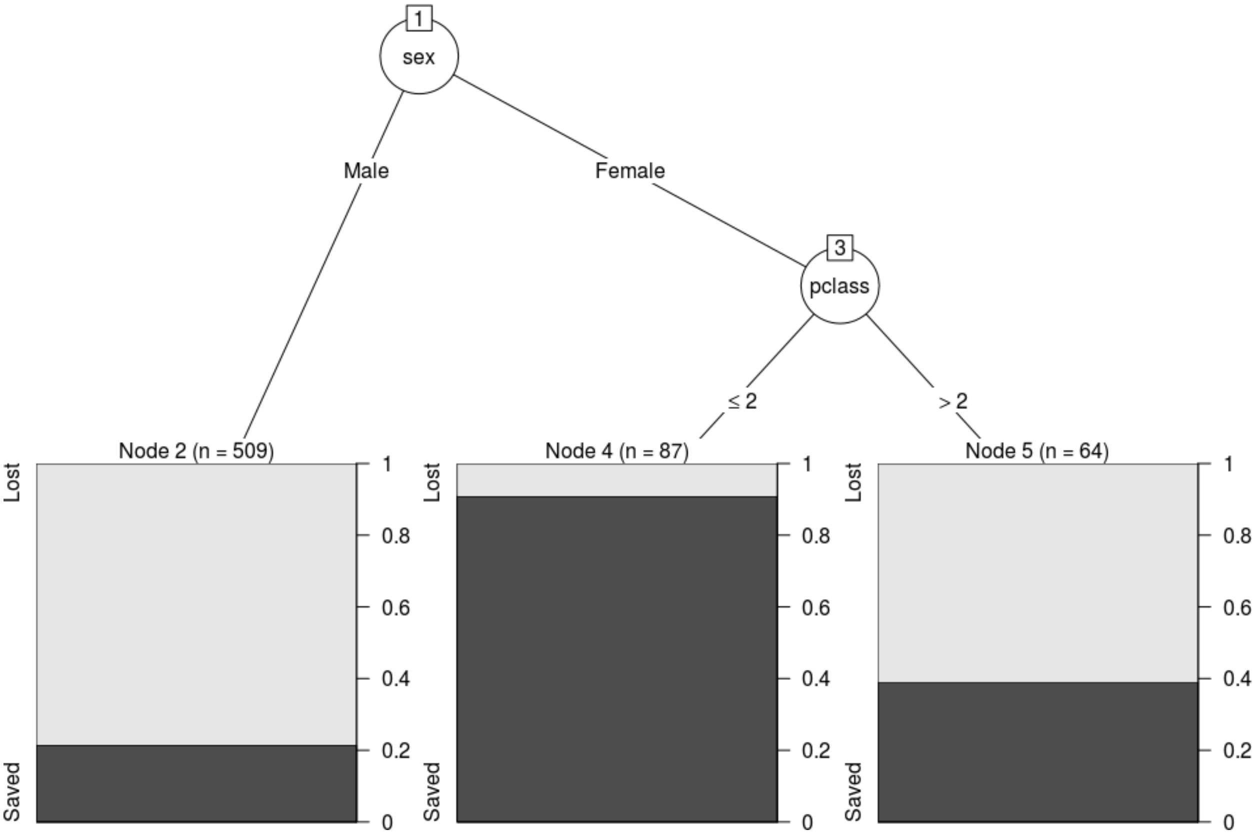

Figure 14: J48 (pruned) decision tree.

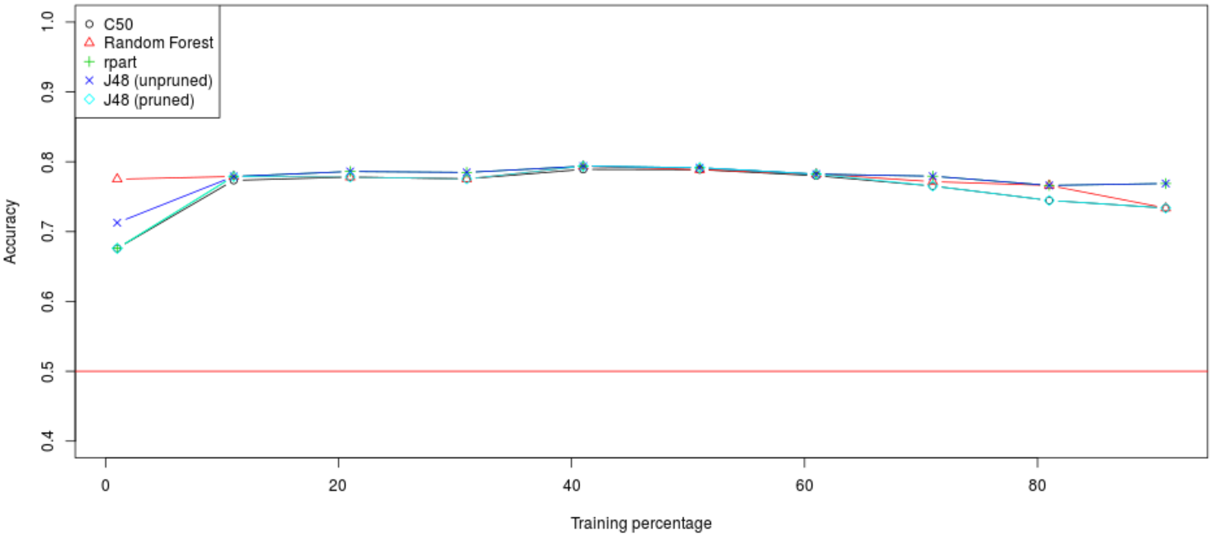

Figure 15: Accuracy based on training dataset size. The horizontal line at 50% represents the accuracy that would be achieved based on using an unbiased coin to decide the likelihood of survival.

A Plotting commands

A collection of commands to create the images in this report.

‘A Mosaic plot of R Titanic data’ on page 5 –

```
library(titanic)
library(graphics)
mosaicplot(Titanic, main = "Survival on the Titanic")
```

‘vcdExtra Titanic data’ on page 7 –

```
library(vcdExtra)
data(Titanicp)
plot(Titanicp)
```

‘vcdExtra Titanic data decision tree based on passenger class, age, and number and type of traveling companions’ on page 8 –

```
library(rpart)
library(rpart.plot)
data(Titanicp, package="vcdExtra")
rp0 <- rpart(survived ~ pclass + age, data=Titanicp)
rpart.plot(rp0, type=0, extra=2, cex=1.5)
```

‘vcdExtra Titanic data decision tree based on all available data’ on page 9 –

```
library(rpart)
library(rpart.plot)
data(Titanicp, package="vcdExtra")
rp0 <- rpart(survived ~ ., data=Titanicp)
rpart.plot(rp0, type=0, extra=2, cex=1.5)
```

‘Missing data map for the “internal” dataset’ on page 11 –

```
source("predictions.R")
library(Amelia)
missmap(d$titanicPeople)
```

‘Missing data map for the “titanic3” dataset’ on page 12 –

```
source("predictions.R")
d <- main(sourceSelection="titanic3")
library(Amelia)
missmap(d$titanicPeople)
```

B Files

A collection of miscellaneous files mentioned in the report.

- titanic3.xls – A consolidated list of passengers. 📎
- predictions.R – An R script to demonstrate the different partitioning approaches and their results. 📎