

SDM4 in R: Displaying and Describing Categorical Data (Chapter 2)

Nicholas Horton (nhorton@amherst.edu) and Sarah McDonald

June 13, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Stats: Data and Models* (2014) by De Veaux, Velleman, and Bock. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at <http://nhorton.people.amherst.edu/sdm4>.

This work leverages initiatives undertaken by Project MOSAIC (<http://www.mosaic-web.org>), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (<http://cran.r-project.org/web/packages/mosaic>). A paper describing the mosaic approach was published in the *R Journal*: <https://journal.r-project.org/archive/2017/RJ-2017-024>.

Chapter 2: Displaying and describing categorical data

Section 2.1: Summarizing and displaying a single categorical variable

See displays on page 19-20.

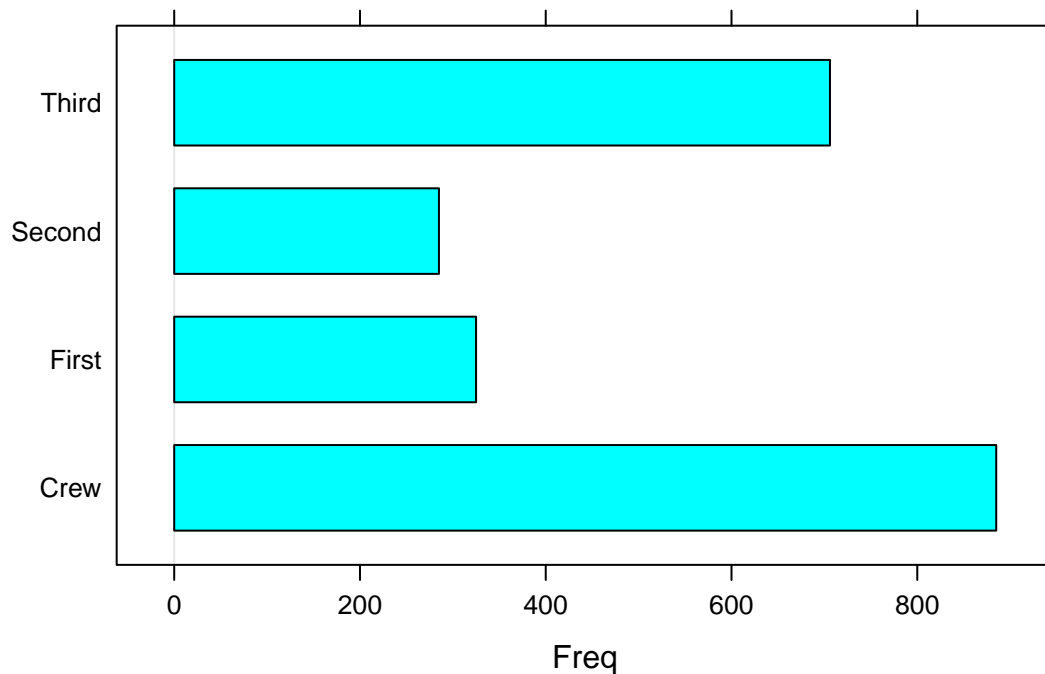
```
library(mosaic)
library(readr)
options(digits = 3)
Titanic <- read_delim("http://nhorton.people.amherst.edu/sdm4/data/Titanic.txt", delim="\t")
tally(~ Class, data = Titanic)
```

```
## Class
##   Crew First Second  Third
##   885   325   285    706
```

```
tally(~ Class, format = "percent", data = Titanic)
```

```
## Class
##   Crew First Second  Third
##   40.2  14.8  12.9   32.1
```

```
barchart(tally(~ Class, data = Titanic))
```



Section 2.2: Exploring the relationship between two categorical variables

See display on page 21.

```
tally(~ Survived + Class, margin = TRUE, data = Titanic)
```

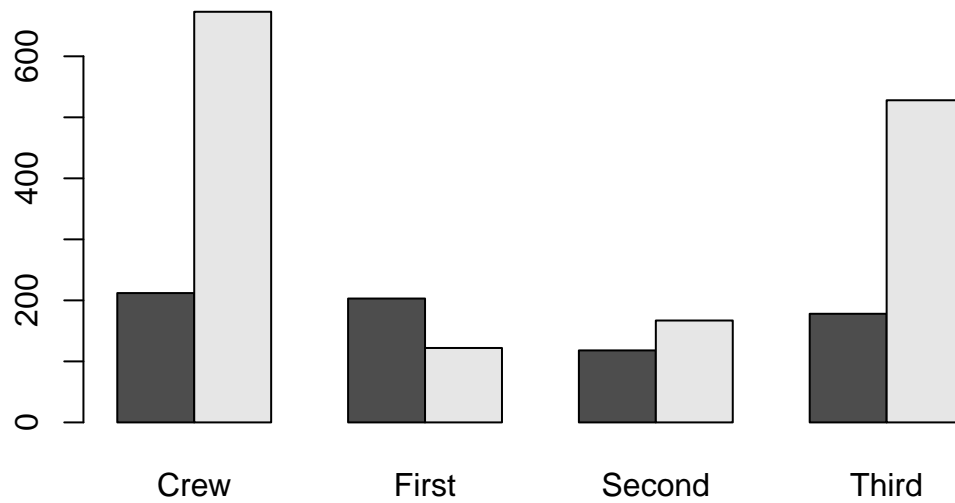
```
##           Class
## Survived Crew First Second Third Total
##   Alive  212   203   118   178   711
##   Dead   673   122   167   528  1490
##   Total  885   325   285   706  2201
```

```
tally(~ Survived | Class, format = "percent", data = Titanic)
```

```
##           Class
## Survived Crew First Second Third
##   Alive 24.0  62.5  41.4  25.2
##   Dead  76.0  37.5  58.6  74.8
```

See display on page 24. (Note that the `mosaic()` plot, also known as Eikosogram, has no connection to the `mosaic` package. We use the `::` notation to access this function from the `vcd` package in R.

```
barplot(tally(~ Survived + Class, data = Titanic), beside = TRUE)
```



```
vcd::mosaic(tally(~ Survived + Class, data = Titanic),
  main = "Mosaic plot of Class by Survival",
  shade = TRUE)
```

Mosaic plot of Class by Survival

