

SDM4 in R: Stats Starts Here (Chapter 1)

Nicholas Horton (nhorton@amherst.edu)

June 12, 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 1: Stats Starts Here

Section 1.1: What is Statistics?

Section 1.2: Data

Section 1.3: Variables

See table on page 7.

```
library(mosaic)
library(readr)
options(digits = 3)
Tour <- read.delim("http://nhorton.people.amherst.edu/sdm4/data/Tour_de_France_2014.txt",
  sep = "\t", stringsAsFactors = FALSE)
names(Tour)
```

```
## [1] "Year"           "Winner"         "Country"
## [4] "Age"           "Team"          "TotalTime.h.min.sec."
## [7] "TotalTime.h." "Average.Speed" "Stages"
## [10] "DistanceRidden" "StartingRiders" "FinishingRiders"
```

```
dim(Tour)
```

```
## [1] 101 12
```

```
head(Tour, 3)
```

```
##   Year      Winner Country Age      Team TotalTime.h.min.sec.
## 1 1903  Maurice Garin  France 32 La Fran\x8daise      94.33.00
```

```
## 2 1904      Henri Cornet  France  20      Cycles JC      96.05.56
## 3 1905 Louis Trousselier  France  24      Peugeot      110.26.58
##   TotalTime.h. Average.Speed Stages DistanceRidden StartingRiders
## 1         94.5         25.7      6         2428         60
## 2         96.1         25.3      6         2428         88
## 3        110.4         27.1     11         2994         60
##   FinishingRiders
## 1             21
## 2             27
## 3             24
```

```
tail(Tour, 8)
```

```
##   Year      Winner      Country Age      Team
## 94 2007  Contador Alberto      Spain 24  Discovery
## 95 2008   Sastre Carlos      Spain 33 CSC-Saxo Bank
## 96 2009  Contador Alberto      Spain 26   Astana
## 97 2010   Andy Schleck  Luxembourg 25   Saxo Bank
## 98 2011   Cadel Evans  Australia 34     BMC
## 99 2012  Bradley Wiggins Great Britain 32     Sky
## 100 2013 Christopher Froome Great Britain 28     Sky
## 101 2014  Vincezo Nibali      Italy 29   Astana
##   TotalTime.h.min.sec. TotalTime.h. Average.Speed Stages DistanceRidden
## 94          91.00.26          91.0         39.2      21         3570
## 95          87.52.52          87.9         40.5      21         3559
## 96          85.48.35          85.8         40.3      21         3460
## 97          91.58.48          92.0         39.6      20         3642
## 98          86.12.22          86.2         39.8      21         3630
## 99          87.34.47          87.6         39.9      20         3497
## 100         94.33.00          94.5         40.5      21         3404
## 101         89.56.06          89.9         40.7      21         3664
##   StartingRiders FinishingRiders
## 94             189             141
## 95             180             145
## 96             180             156
## 97             198             170
## 98             198             167
## 99             198             153
## 100            198             169
## 101            198             164
```

Let's find who was the winner in 1998

```
filter(Tour, Year == 1998)
```

```
##   Year      Winner Country Age      Team TotalTime.h.min.sec.
## 1 1998 Marco Pantani  Italy 28 Mercatone Uno      92.49.46
##   TotalTime.h. Average.Speed Stages DistanceRidden StartingRiders
## 1         92.8         40      21         3875         189
##   FinishingRiders
## 1             96
```

How many stages were there in the Tour in the years that Alberto Contador won?

```
filter(Tour, Winner == "Contador Alberto")
```

```
##   Year      Winner Country Age      Team TotalTime.h.min.sec.
## 1 2007 Contador Alberto  Spain 24 Discovery          91.00.26
## 2 2009 Contador Alberto  Spain 26  Astana           85.48.35
##   TotalTime.h. Average.Speed Stages DistanceRidden StartingRiders
## 1          91.0          39.2    21           3570           189
## 2          85.8          40.3    21           3460           180
##   FinishingRiders
## 1                141
## 2                156
```

Note that the following commands generate the same output.

```
Tour %>%
  filter(Winner == "Contador Alberto")
```

```
##   Year      Winner Country Age      Team TotalTime.h.min.sec.
## 1 2007 Contador Alberto  Spain 24 Discovery          91.00.26
## 2 2009 Contador Alberto  Spain 26  Astana           85.48.35
##   TotalTime.h. Average.Speed Stages DistanceRidden StartingRiders
## 1          91.0          39.2    21           3570           189
## 2          85.8          40.3    21           3460           180
##   FinishingRiders
## 1                141
## 2                156
```

The pipe operator (%>%) can be used to connect one dataframe or command to another.

What was the slowest average speed of any tour? Fastest?

```
filter(Tour, Average.Speed == min(Average.Speed))
```

```
##   Year      Winner Country Age      Team TotalTime.h.min.sec.
## 1 1919 Fir Lambot Belgium 33 La Sportive          231.07.15
##   TotalTime.h. Average.Speed Stages DistanceRidden StartingRiders
## 1          231          24.1    15           5560           69
##   FinishingRiders
## 1                11
```

```
filter(Tour, Average.Speed == max(Average.Speed))
```

```
##   Year      Winner Country Age      Team TotalTime.h.min.sec.
## 1 2005 Lance Armstrong  USA 34 Discovery          86.15.02
##   TotalTime.h. Average.Speed Stages DistanceRidden StartingRiders
## 1          86.3          41.7    21           3593           189
##   FinishingRiders
## 1                155
```

What can we say about the Average Speeds?

```
favstats(~ Average.Speed, data = Tour)
```

```
##   min   Q1 median   Q3  max mean   sd   n missing  
## 24.1 29.1  35.4 38.6 41.7   34 5.19 101      0
```