# Amazon Books Analysis

Haochong(Rogers) Yang

04/08/2022

**This file mainly examined books sold by Amazon, in which I queried some detail information, such as the minimum pages of book and summary tables.**

The code below reads in data about books sold on Amazon (https://dasl.datadescription.com/datafile/amazon-books/). The data frame containing the book data is named `books`. Note that the height (`Height`), width (`Width`) and thickness (`Thick`) of books in this data frame are measured in inches.

```
books <- read.csv("amazonbooks.csv")
```

**What is the name of the book(s) with the smallest number of pages in this sample of books, and how many pages does it have?**

```
books %>% arrange(NumPages) %>% select(Title, NumPages) %>% head()
```

```
##                                                                Title NumPages
## 1                                           Big Dog . . . Little Dog       24
## 2                             The Berenstain Bears He Bear, She Bear       24
## 3 The Shape of Me and Other Stuff: Dr. Seuss’s Surprising Word Book       24
## 4                                  Cloudy With a Chance of Meatballs       32
## 5                                                 Go the F**k Asleep       32
## 6                                                           Madeline       54
```

There are three books that have the same lowest number of pages, which is 24 pages. The names are “Big Dog . . . Little Dog”, “The Berenstain Bears He Bear, She Bear”, and “The Shape of Me and Other Stuff: Dr. Seuss’s Surprising Word Book”.

**(b) I create a summary table which reports the total number of books written by each author and the mean and variance of the number of pages per book for each author, for the books represented in this sample of books.**

```
books %>% group_by(Author) %>%
  summarize(n = n(),
            mean_pages = mean(NumPages),
            var_pages = var(NumPages))
```

```
## # A tibble: 256 x 4
##    Author                 n mean_pages var_pages
##    <chr>              <int>      <dbl>     <dbl>
##  1 ""                     1        432        NA
##  2 "Abraham Verghese"     1        667        NA
##  3 "Adam Goodheart"       1        460        NA
##  4 "Adam Hochschild"      1        480        NA
##  5 "Adam Mansbach"        1         32        NA
##  6 "Alaa Aswany"          1        255        NA
##  7 "Alice Munro"          2        320      2048
##  8 "Alice Schroeder"      1        832        NA
##  9 "Allen, Toorawa"       1        200        NA
## 10 "Andrea Warren"        1        160        NA
## # ... with 246 more rows
```

**(c) I created a new summary table based on the previous one which contains only information for authors who wrote more than 2 books, and sorted them in decreasing order of number of books written.**

```
books %>% group_by(Author) %>%
  summarize(n = n(), mean_pages = mean(NumPages), var_pages = var(NumPages)) %>%
  filter(n > 1) %>% arrange(desc(n))
```

```
## # A tibble: 43 x 4
##    Author               n mean_pages var_pages
##    <chr>            <int>      <dbl>     <dbl>
##  1 Jodi Picoult         7       414.     1658.
##  2 Vladimir Nabokov     7       316     20528
##  3 Lewis                4       266.    18820.
##  4 Murakami             4       354.     9838.
##  5 Ben Mezrich          3       299        571
##  6 Bruce Ballenger      3       448       9472
##  7 Christensen          3       245.    24917.
##  8 Collins              3       370.     1920.
##  9 Drucker              3       304     11008
## 10 Ha Jin               3       300      5232
## # ... with 33 more rows
```