Today, we'll start digging into some of the functions used to summarise data. The full summarise function will be covered for the letter S. For now, let's look at one function from the tidyverse that can give some overall information about a dataset: n_distinct.

This function counts the number of unique values in a vector or variable. There are 87 books in my 2019 reading list, but I read multiple books by the same author(s). Let's see how many authors there are in my set.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse
1.3.0 --
```

```
##  ggplot2 3.2.1      purrr   0.3.3
##  tibble  2.1.3      dplyr   0.8.3
##  tidyr   1.0.0      stringr 1.4.0
##  readr   1.3.1      forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
reads2019 <- read_csv("~/Downloads/Blogging A to Z/SaraReads2019_allrated.csv",
col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   Title = col_character(),
##   Pages = col_double(),
##   date_started = col_character(),
##   date_read = col_character(),
##   Book.ID = col_double(),
##   Author = col_character(),
##   AdditionalAuthors = col_character(),
##   AverageRating = col_double(),
##   OriginalPublicationYear = col_double(),
##   read_time = col_double(),
##   MyRating = col_double(),
##   Gender = col_double(),
##   Fiction = col_double(),
##   Childrens = col_double(),
##   Fantasy = col_double(),
##   SciFi = col_double(),
##   Mystery = col_double(),
##   SelfHelp = col_double()
## )
```

```
reads2019 %$% n_distinct(Author)
```

```
## [1] 42
```

So while there are 87 books in my dataset, there are only 42 authors. Let's see who the top authors are.

```
reads2019 %>%
  group_by(Author) %>%
  summarise(Books = n()) %>%
  arrange(desc(Books), Author) %>%
  filter(between(row_number(), 1, 10))
```

```
## # A tibble: 10 x 2
##    Author           Books
##
##  1 Baum, L. Frank      14
##  2 Pratchett, Terry    13
##  3 King, Stephen        6
##  4 Scalzi, John         6
##  5 Abbott, Mildred      5
##  6 Atwood, Margaret     5
##  7 Patchett, Ann        2
##  8 Ware, Ruth           2
##  9 Adams, Douglas       1
## 10 Adeyemi, Tomi        1
```

14 books were written by L. Frank Baum – this makes sense, because one of my goals was to reread the Oz book series, of which there are 14, starting with *The Wonderful Wizard of Oz* and ending with *Glinda of Oz*. 13 are by Terry Pratchett (mostly books from the Discworld series). And finally, Stephen King and John Scalzi are tied for 3rd, with 6 books each.

n_distinct can also be used in conjunction with other functions, like filter or group_by.

```
library(tidytext)

titlewords <- reads2019 %>%
  unnest_tokens(titleword, Title) %>%
  select(titleword, Author, Book.ID) %>%
  left_join(reads2019, by = c("Book.ID", "Author"))

titlewords %>%
  group_by(Title) %>%
  summarise(unique_words = n_distinct(titleword),
            total_words = n())
```

```
## # A tibble: 87 x 3
##    Title                                                 unique_words
total_words
##
##  1 1Q84                                                             1
1
##  2 A Disorder Peculiar to the Country                               6
6
##  3 Alas, Babylon                                                    2
2
##  4 Artemis                                                          1
1
##  5 Bird Box (Bird Box, #1)                                          3
5
##  6 Boundaries: When to Say Yes, How to Say No to Take …            12
```

```
15
##  7 Cell                                                                1
1
##  8 Children of Virtue and Vengeance (Legacy of Orïsha,…                8
9
##  9 Cujo                                                                1
1
## 10 Dirk Gently's Holistic Detective Agency (Dirk Gentl…                7
8
## # … with 77 more rows
```

This chunk of code separated title into its individual words, then counted the number of unique words within each book title. For many cases, some words are reused multiple times in the same title – often words like “of” or “to”. We could also write some code to tell us how many unique words are used across all titles.

```
titlewords %$%
  n_distinct(titleword)
```

```
## [1] 224
```

There are, overall, 458 titlewords that make up the titles of the books in the dataset, but only 224 distinct words are used. This means that many titles are using the same words as others. Once again, these are probably common words. Let’s see what happens when we remove those common words.

```
titlewords <- titlewords %>%
  anti_join(stop_words, by = c("titleword" = "word"))

titlewords %$%
  n_distinct(titleword)
```

```
## [1] 181
```

After removing stopwords, there are now 306 individual words, but only 181 distinct ones.