Let's now continue to explore the tidyverse with the `dplyr` package. Here we'll learn about five verbs and the pipe operator.

```
library(dplyr) # for manipulating data
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr) # for getting data
library(ggplot2) # for plotting data

tips <- read_csv("tips.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   total_bill = col_double(),
##   tip = col_double(),
##   sex = col_character(),
##   smoker = col_character(),
##   day = col_character(),
##   time = col_character(),
##   size = col_double()
## )
```

```
# You can change where this is output above in "Settings (by knit) > Chunk Output in Console"
tips
```

```
## # A tibble: 244 x 8
##       X1 total_bill   tip sex    smoker day   time    size
##    <dbl>      <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>
##  1     0      17.0   1.01 Female No     Sun   Dinner     2
##  2     1      10.3   1.66 Male   No     Sun   Dinner     3
##  3     2      21.0   3.5  Male   No     Sun   Dinner     3
##  4     3      23.7   3.31 Male   No     Sun   Dinner     2
##  5     4      24.6   3.61 Female No     Sun   Dinner     4
##  6     5      25.3   4.71 Male   No     Sun   Dinner     4
##  7     6       8.77  2    Male   No     Sun   Dinner     2
##  8     7      26.9   3.12 Male   No     Sun   Dinner     4
##  9     8      15.0   1.96 Male   No     Sun   Dinner     2
## 10     9      14.8   3.23 Male   No     Sun   Dinner     2
## # ... with 234 more rows
```

The first verb we will use is `select()` which lets us choose colunmns.

```
select(tips, total_bill)
```

```
## # A tibble: 244 x 1
##    total_bill
##         <dbl>
##  1      17.0
##  2      10.3
##  3      21.0
##  4      23.7
##  5      24.6
##  6      25.3
##  7       8.77
##  8      26.9
##  9      15.0
## 10      14.8
## # ... with 234 more rows
```

```
select(tips, -X1)
```

```
## # A tibble: 244 x 7
##    total_bill   tip sex    smoker day   time    size
##         <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>
##  1      17.0   1.01 Female No     Sun   Dinner     2
##  2      10.3   1.66 Male   No     Sun   Dinner     3
##  3      21.0   3.5  Male   No     Sun   Dinner     3
##  4      23.7   3.31 Male   No     Sun   Dinner     2
##  5      24.6   3.61 Female No     Sun   Dinner     4
##  6      25.3   4.71 Male   No     Sun   Dinner     4
##  7       8.77  2    Male   No     Sun   Dinner     2
##  8      26.9   3.12 Male   No     Sun   Dinner     4
##  9      15.0   1.96 Male   No     Sun   Dinner     2
## 10      14.8   3.23 Male   No     Sun   Dinner     2
## # ... with 234 more rows
```

```
select(tips, tip:size)
```

```
## # A tibble: 244 x 6
##      tip sex    smoker day   time    size
##    <dbl> <chr>  <chr>  <chr> <chr>  <dbl>
##  1  1.01 Female No     Sun   Dinner     2
##  2  1.66 Male   No     Sun   Dinner     3
##  3  3.5  Male   No     Sun   Dinner     3
##  4  3.31 Male   No     Sun   Dinner     2
##  5  3.61 Female No     Sun   Dinner     4
##  6  4.71 Male   No     Sun   Dinner     4
##  7  2    Male   No     Sun   Dinner     2
##  8  3.12 Male   No     Sun   Dinner     4
##  9  1.96 Male   No     Sun   Dinner     2
## 10  3.23 Male   No     Sun   Dinner     2
## # ... with 234 more rows
```

```
select(tips, starts_with("s"))
```

```
## # A tibble: 244 x 3
##    sex    smoker  size
##    <chr>  <chr>  <dbl>
##  1 Female No         2
##  2 Male   No         3
##  3 Male   No         3
##  4 Male   No         2
##  5 Female No         4
##  6 Male   No         4
##  7 Male   No         2
##  8 Male   No         4
##  9 Male   No         2
## 10 Male   No         2
## # ... with 234 more rows
```

Now let's check out filter!

```
filter(tips, day == "Sun")
```

```
## # A tibble: 76 x 8
##       X1 total_bill   tip sex    smoker day   time    size
##    <dbl>      <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>
##  1     0      17.0   1.01 Female No     Sun   Dinner     2
##  2     1      10.3   1.66 Male   No     Sun   Dinner     3
##  3     2      21.0   3.5  Male   No     Sun   Dinner     3
##  4     3      23.7   3.31 Male   No     Sun   Dinner     2
##  5     4      24.6   3.61 Female No     Sun   Dinner     4
##  6     5      25.3   4.71 Male   No     Sun   Dinner     4
##  7     6       8.77  2    Male   No     Sun   Dinner     2
##  8     7      26.9   3.12 Male   No     Sun   Dinner     4
##  9     8      15.0   1.96 Male   No     Sun   Dinner     2
## 10     9      14.8   3.23 Male   No     Sun   Dinner     2
## # ... with 66 more rows
```

```
filter(tips, tip > 5)
```

```
## # A tibble: 18 x 8
##       X1 total_bill   tip sex    smoker day   time    size
##    <dbl>      <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>
##  1    23      39.4   7.58 Male   No     Sat   Dinner     4
##  2    44      30.4   5.6  Male   No     Sun   Dinner     4
##  3    47      32.4   6    Male   No     Sun   Dinner     4
##  4    52      34.8   5.2  Female No     Sun   Dinner     4
##  5    59      48.3   6.73 Male   No     Sat   Dinner     4
##  6    85      34.8   5.17 Female No     Thur  Lunch      4
##  7    88      24.7   5.85 Male   No     Thur  Lunch      2
##  8   116      29.9   5.07 Male   No     Sun   Dinner     4
##  9   141      34.3   6.7  Male   No     Thur  Lunch      6
## 10   155      29.8   5.14 Female No     Sun   Dinner     5
```

```
## 11   170      50.8  10    Male   Yes    Sat   Dinner     3
## 12   172       7.25  5.15 Male   Yes    Sun   Dinner     2
## 13   181      23.3   5.65 Male   Yes    Sun   Dinner     2
## 14   183      23.2   6.5  Male   Yes    Sun   Dinner     4
## 15   211      25.9   5.16 Male   Yes    Sat   Dinner     4
## 16   212      48.3   9    Male   No     Sat   Dinner     4
## 17   214      28.2   6.5  Female Yes    Sat   Dinner     3
## 18   239      29.0   5.92 Male   No     Sat   Dinner     3
```

```
filter(tips, sex == "Male" & smoker == "Yes")
```

```
## # A tibble: 60 x 8
##       X1 total_bill   tip sex   smoker day   time    size
##    <dbl>      <dbl> <dbl> <chr> <chr>  <chr> <chr>  <dbl>
##  1    56       38.0  3    Male  Yes    Sat   Dinner     4
##  2    58       11.2  1.76 Male  Yes    Sat   Dinner     2
##  3    60       20.3  3.21 Male  Yes    Sat   Dinner     2
##  4    61       13.8  2    Male  Yes    Sat   Dinner     2
##  5    62       11.0  1.98 Male  Yes    Sat   Dinner     2
##  6    63       18.3  3.76 Male  Yes    Sat   Dinner     4
##  7    69       15.0  2.09 Male  Yes    Sat   Dinner     2
##  8    76       17.9  3.08 Male  Yes    Sat   Dinner     2
##  9    80       19.4  3    Male  Yes    Thur  Lunch      2
## 10    83       32.7  5    Male  Yes    Thur  Lunch      2
## # ... with 50 more rows
```

```
filter(tips, sex == "Male" | smoker == "Yes")
```

```
## # A tibble: 190 x 8
##       X1 total_bill   tip sex   smoker day   time    size
##    <dbl>      <dbl> <dbl> <chr> <chr>  <chr> <chr>  <dbl>
##  1     1      10.3   1.66 Male  No     Sun   Dinner     3
##  2     2      21.0   3.5  Male  No     Sun   Dinner     3
##  3     3      23.7   3.31 Male  No     Sun   Dinner     2
##  4     5      25.3   4.71 Male  No     Sun   Dinner     4
##  5     6       8.77  2    Male  No     Sun   Dinner     2
##  6     7      26.9   3.12 Male  No     Sun   Dinner     4
##  7     8      15.0   1.96 Male  No     Sun   Dinner     2
##  8     9      14.8   3.23 Male  No     Sun   Dinner     2
##  9    10      10.3   1.71 Male  No     Sun   Dinner     2
## 10    12      15.4   1.57 Male  No     Sun   Dinner     2
## # ... with 180 more rows
```

Let's now return to our example from before to see `mutate()` in action. It's the same output, but right now looks a bit different since it's simplified.

```
mutate(tips, gbp_total = total_bill * 0.81)
```

```
## # A tibble: 244 x 9
##       X1 total_bill   tip sex    smoker day   time    size gbp_total
##    <dbl>      <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>     <dbl>
##  1     0      17.0   1.01 Female No     Sun   Dinner     2     13.8
```

```
##  2     1      10.3   1.66 Male   No     Sun   Dinner     3      8.38
##  3     2      21.0   3.5  Male   No     Sun   Dinner     3     17.0
##  4     3      23.7   3.31 Male   No     Sun   Dinner     2     19.2
##  5     4      24.6   3.61 Female No     Sun   Dinner     4     19.9
##  6     5      25.3   4.71 Male   No     Sun   Dinner     4     20.5
##  7     6       8.77  2    Male   No     Sun   Dinner     2      7.10
##  8     7      26.9   3.12 Male   No     Sun   Dinner     4     21.8
##  9     8      15.0   1.96 Male   No     Sun   Dinner     2     12.2
## 10     9      14.8   3.23 Male   No     Sun   Dinner     2     12.0
## # ... with 234 more rows
```

```
# R doesn't care about spacing!!
mutate(tips,
       gbp_total = total_bill * 0.81,
       gbp_tip = tip * 0.81)
```

```
## # A tibble: 244 x 10
##       X1 total_bill   tip sex    smoker day   time    size gbp_total gbp_tip
##    <dbl>      <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>     <dbl>   <dbl>
##  1     0      17.0   1.01 Female No     Sun   Dinner     2     13.8    0.818
##  2     1      10.3   1.66 Male   No     Sun   Dinner     3      8.38   1.34
##  3     2      21.0   3.5  Male   No     Sun   Dinner     3     17.0    2.84
##  4     3      23.7   3.31 Male   No     Sun   Dinner     2     19.2    2.68
##  5     4      24.6   3.61 Female No     Sun   Dinner     4     19.9    2.92
##  6     5      25.3   4.71 Male   No     Sun   Dinner     4     20.5    3.82
##  7     6       8.77  2    Male   No     Sun   Dinner     2      7.10   1.62
##  8     7      26.9   3.12 Male   No     Sun   Dinner     4     21.8    2.53
##  9     8      15.0   1.96 Male   No     Sun   Dinner     2     12.2    1.59
## 10     9      14.8   3.23 Male   No     Sun   Dinner     2     12.0    2.62
## # ... with 234 more rows
```

Up until this point, we have been doing some very basic commands. What if we want to get super fancy? We could save a bunch of intermediate objects into memory... or we could use the %>% operator!

```
select(tips, total_bill)
```

```
## # A tibble: 244 x 1
##    total_bill
##         <dbl>
##  1      17.0
##  2      10.3
##  3      21.0
##  4      23.7
##  5      24.6
##  6      25.3
##  7       8.77
##  8      26.9
##  9      15.0
## 10      14.8
## # ... with 234 more rows
```

```
tips %>%  # and then!
  select(total_bill)
```

```
## # A tibble: 244 x 1
##    total_bill
##         <dbl>
##  1      17.0
##  2      10.3
##  3      21.0
##  4      23.7
##  5      24.6
##  6      25.3
##  7       8.77
##  8      26.9
##  9      15.0
## 10      14.8
## # ... with 234 more rows
```

We can then build up more complex commands.

```
tips %>%
  select(total_bill, tip, sex, smoker) %>%
  filter(sex == "Male" & smoker == "Yes") %>%
  mutate(gbp_total_bill = total_bill * 0.81,
         gbp_tip = tip * 0.81)
```

```
## # A tibble: 60 x 6
##    total_bill   tip sex   smoker gbp_total_bill gbp_tip
##         <dbl> <dbl> <chr> <chr>           <dbl>   <dbl>
##  1       38.0  3    Male  Yes             30.8     2.43
##  2       11.2  1.76 Male  Yes              9.10    1.43
##  3       20.3  3.21 Male  Yes             16.4     2.60
##  4       13.8  2    Male  Yes             11.2     1.62
##  5       11.0  1.98 Male  Yes              8.93    1.60
##  6       18.3  3.76 Male  Yes             14.8     3.05
##  7       15.0  2.09 Male  Yes             12.2     1.69
##  8       17.9  3.08 Male  Yes             14.5     2.49
##  9       19.4  3    Male  Yes             15.7     2.43
## 10       32.7  5    Male  Yes             26.5     4.05
## # ... with 50 more rows
```

But what if we wanted to investigate differences between smoker and non smokers for tipping? For that we need group_by and summerise!

```
tips %>%
  group_by(smoker)
```

```
## # A tibble: 244 x 8
## # Groups:   smoker [2]
##       X1 total_bill   tip sex    smoker day   time    size
##    <dbl>      <dbl> <dbl> <chr>  <chr>  <chr> <chr>  <dbl>
```

```
##  1     0      17.0   1.01 Female No     Sun   Dinner     2
##  2     1      10.3   1.66 Male   No     Sun   Dinner     3
##  3     2      21.0   3.5  Male   No     Sun   Dinner     3
##  4     3      23.7   3.31 Male   No     Sun   Dinner     2
##  5     4      24.6   3.61 Female No     Sun   Dinner     4
##  6     5      25.3   4.71 Male   No     Sun   Dinner     4
##  7     6       8.77  2    Male   No     Sun   Dinner     2
##  8     7      26.9   3.12 Male   No     Sun   Dinner     4
##  9     8      15.0   1.96 Male   No     Sun   Dinner     2
## 10     9      14.8   3.23 Male   No     Sun   Dinner     2
## # ... with 234 more rows
```

```
tips %>%
  group_by(smoker) %>%
  summarise(mean = mean(tip),
            count = n())
```

```
## # A tibble: 2 x 3
##   smoker  mean count
##   <chr>  <dbl> <int>
## 1 No      2.99   151
## 2 Yes     3.01    93
```

Lastly let's arrange our output here so that those who tip more are on top!

```
tips %>%
  group_by(smoker) %>%
  summarise(mean = mean(tip),
            count = n()) %>%
  arrange(desc(mean))
```

```
## # A tibble: 2 x 3
##   smoker  mean count
##   <chr>  <dbl> <int>
## 1 Yes     3.01    93
## 2 No      2.99   151
```