tidyverse and data.table

```
library(pacman)
p_load(nycflights13, tidyverse)
```

Tidyverse: dplyr and tidyr

dplyr

1. filter

Filter (i.e. subset) rows based on their values.

```
starwars %>%
  filter(
    species == "Human",
    height >= 190
    )
```

```
## # A tibble: 4 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
## 2 Qui-Gon ~    193    89 brown      fair       blue            92   male  mascu~
## 3 Dooku        193    80 white      fair       brown          102   male  mascu~
## 4 Bail Pre~    191    NA black      tan        brown           67   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
starwars %>%
  filter(grepl("Skywalker", name))
```

```
## # A tibble: 3 x 14
##   name     height  mass hair_color skin_color eye_color birth_year sex    gender
##   <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>  <chr>
## 1 Luke Sk~    172    77 blond      fair       blue            19   male   mascu~
## 2 Anakin ~    188    84 blond      fair       blue            41.9 male   mascu~
## 3 Shmi Sk~    163    NA black      fair       brown           72   female femin~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
starwars %>%
  filter(is.na(height))
```

```
## # A tibble: 6 x 14
##   name     height  mass hair_color skin_color eye_color birth_year sex    gender
```

```
##   <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>  <chr>
## 1 Arvel C~     NA    NA brown      fair       brown             NA male   mascu~
## 2 Finn         NA    NA black      dark       dark              NA male   mascu~
## 3 Rey          NA    NA brown      light      hazel             NA female femin~
## 4 Poe Dam~     NA    NA brown      light      brown             NA male   mascu~
## 5 BB8          NA    NA none       none       black             NA none   mascu~
## 6 Captain~     NA    NA unknown    unknown    unknown           NA <NA>   <NA>
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
starwars %>%
  filter(!is.na(height))
```

```
## # A tibble: 81 x 14
##    name    height  mass hair_color  skin_color eye_color birth_year sex   gender
##    <chr>    <int> <dbl> <chr>       <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke S~    172    77 blond       fair       blue            19   male  mascu~
##  2 C-3PO      167    75 <NA>        gold       yellow         112   none  mascu~
##  3 R2-D2       96    32 <NA>        white, bl~ red             33   none  mascu~
##  4 Darth ~    202   136 none        white      yellow          41.9 male  mascu~
##  5 Leia O~    150    49 brown       light      brown           19   fema~ femin~
##  6 Owen L~    178   120 brown, grey light      blue            52   male  mascu~
##  7 Beru W~    165    75 brown       light      blue            47   fema~ femin~
##  8 R5-D4       97    32 <NA>        white, red red             NA   none  mascu~
##  9 Biggs ~    183    84 black       light      brown           24   male  mascu~
## 10 Obi-Wa~    182    77 auburn, wh~ fair       blue-gray       57   male  mascu~
## # ... with 71 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

2. arrange

Arrange (i.e. reorder) rows based on their values.

```
starwars %>%
  arrange(birth_year)
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Wicket ~     88  20   brown      brown      brown            8   male  mascu~
##  2 IG-88       200 140   none       metal      red             15   none  mascu~
##  3 Luke Sk~    172  77   blond      fair       blue            19   male  mascu~
##  4 Leia Or~    150  49   brown      light      brown           19   fema~ femin~
##  5 Wedge A~    170  77   brown      fair       hazel           21   male  mascu~
##  6 Plo Koon    188  80   none       orange     black           22   male  mascu~
##  7 Biggs D~    183  84   black      light      brown           24   male  mascu~
##  8 Han Solo    180  80   brown      fair       brown           29   male  mascu~
##  9 Lando C~    177  79   black      dark       brown           31   male  mascu~
## 10 Boba Fe~    183  78.2 black      fair       brown           31.5 male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
starwars %>%
  arrange(desc(birth_year))
```

```
## # A tibble: 87 x 14
##    name    height  mass hair_color skin_color  eye_color birth_year sex   gender
##    <chr>    <int> <dbl> <chr>      <chr>       <chr>           <dbl> <chr> <chr>
##  1 Yoda        66    17 white      green       brown             896 male  mascu~
##  2 Jabba ~    175  1358 <NA>       green-tan,~ orange            600 herm~ mascu~
##  3 Chewba~    228   112 brown      unknown     blue              200 male  mascu~
##  4 C-3PO      167    75 <NA>       gold        yellow            112 none  mascu~
##  5 Dooku      193    80 white      fair        brown             102 male  mascu~
##  6 Qui-Go~    193    89 brown      fair        blue               92 male  mascu~
##  7 Ki-Adi~    198    82 white      pale        yellow             92 male  mascu~
##  8 Finis ~    170    NA blond      fair        blue               91 male  mascu~
##  9 Palpat~    170    75 grey       pale        yellow             82 male  mascu~
## 10 Cliegg~    183    NA brown      fair        blue               82 male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

3. select

Select (i.e. subset) columns by their names:

```
starwars %>%
  select(name:skin_color, species, -height)
```

```
## # A tibble: 87 x 5
##    name                mass hair_color    skin_color  species
##    <chr>              <dbl> <chr>         <chr>       <chr>
##  1 Luke Skywalker        77 blond         fair        Human
##  2 C-3PO                 75 <NA>          gold        Droid
##  3 R2-D2                 32 <NA>          white, blue Droid
##  4 Darth Vader          136 none          white       Human
##  5 Leia Organa           49 brown         light       Human
##  6 Owen Lars            120 brown, grey   light       Human
##  7 Beru Whitesun lars    75 brown         light       Human
##  8 R5-D4                 32 <NA>          white, red  Droid
##  9 Biggs Darklighter     84 black         light       Human
## 10 Obi-Wan Kenobi        77 auburn, white fair        Human
## # ... with 77 more rows
```

```
starwars %>%
  select(alias=name, crib=homeworld, sex=gender)
```

```
## # A tibble: 87 x 3
##    alias              crib     sex
##    <chr>              <chr>    <chr>
##  1 Luke Skywalker     Tatooine masculine
##  2 C-3PO              Tatooine masculine
##  3 R2-D2              Naboo    masculine
##  4 Darth Vader        Tatooine masculine
```

```
##  5 Leia Organa        Alderaan feminine
##  6 Owen Lars          Tatooine masculine
##  7 Beru Whitesun lars Tatooine feminine
##  8 R5-D4              Tatooine masculine
##  9 Biggs Darklighter  Tatooine masculine
## 10 Obi-Wan Kenobi     Stewjon  masculine
## # ... with 77 more rows
```

```
starwars %>%
  select(name, contains("color"))
```

```
## # A tibble: 87 x 4
##    name               hair_color    skin_color  eye_color
##    <chr>              <chr>         <chr>       <chr>
##  1 Luke Skywalker     blond         fair        blue
##  2 C-3PO              <NA>          gold        yellow
##  3 R2-D2              <NA>          white, blue red
##  4 Darth Vader        none          white       yellow
##  5 Leia Organa        brown         light       brown
##  6 Owen Lars          brown, grey   light       blue
##  7 Beru Whitesun lars brown         light       blue
##  8 R5-D4              <NA>          white, red  red
##  9 Biggs Darklighter  black         light       brown
## 10 Obi-Wan Kenobi     auburn, white fair        blue-gray
## # ... with 77 more rows
```

The select(..., everything()) option is another useful shortcut if you only want to bring some variable(s) to the "front" of a data frame.

```
starwars %>%
  select(species, homeworld, everything()) %>%
  head(5)
```

```
## # A tibble: 5 x 14
##   species homeworld name           height  mass hair_color skin_color  eye_color
##   <chr>   <chr>     <chr>           <int> <dbl> <chr>      <chr>       <chr>
## 1 Human   Tatooine  Luke Skywalker    172    77 blond      fair        blue
## 2 Droid   Tatooine  C-3PO             167    75 <NA>       gold        yellow
## 3 Droid   Naboo     R2-D2              96    32 <NA>       white, blue red
## 4 Human   Tatooine  Darth Vader       202   136 none       white       yellow
## 5 Human   Alderaan  Leia Organa       150    49 brown      light       brown
## # ... with 6 more variables: birth_year <dbl>, sex <chr>, gender <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

4. mutate

Create new columns.

```
starwars %>%
  select(name, birth_year) %>%
  mutate(dog_years = birth_year * 7) %>%
```

```
  mutate(comment = paste0(name, " is ", dog_years, " in dog years."))
```

```
## # A tibble: 87 x 4
##    name               birth_year dog_years comment
##    <chr>                   <dbl>     <dbl> <chr>
##  1 Luke Skywalker           19        133  Luke Skywalker is 133 in dog years.
##  2 C-3PO                   112        784  C-3PO is 784 in dog years.
##  3 R2-D2                    33        231  R2-D2 is 231 in dog years.
##  4 Darth Vader              41.9      293. Darth Vader is 293.3 in dog years.
##  5 Leia Organa              19        133  Leia Organa is 133 in dog years.
##  6 Owen Lars                52        364  Owen Lars is 364 in dog years.
##  7 Beru Whitesun lars       47        329  Beru Whitesun lars is 329 in dog yea~
##  8 R5-D4                    NA         NA  R5-D4 is NA in dog years.
##  9 Biggs Darklighter        24        168  Biggs Darklighter is 168 in dog year~
## 10 Obi-Wan Kenobi           57        399  Obi-Wan Kenobi is 399 in dog years.
## # ... with 77 more rows
```

same as the last one. We can chain multiple mutates in a single call.

```
starwars %>%
  select(name, birth_year) %>%
  mutate(
    dog_years = birth_year * 7, ## Separate with a comma
    comment = paste0(name, " is ", dog_years, " in dog years.")
    )
```

```
## # A tibble: 87 x 4
##    name               birth_year dog_years comment
##    <chr>                   <dbl>     <dbl> <chr>
##  1 Luke Skywalker           19        133  Luke Skywalker is 133 in dog years.
##  2 C-3PO                   112        784  C-3PO is 784 in dog years.
##  3 R2-D2                    33        231  R2-D2 is 231 in dog years.
##  4 Darth Vader              41.9      293. Darth Vader is 293.3 in dog years.
##  5 Leia Organa              19        133  Leia Organa is 133 in dog years.
##  6 Owen Lars                52        364  Owen Lars is 364 in dog years.
##  7 Beru Whitesun lars       47        329  Beru Whitesun lars is 329 in dog yea~
##  8 R5-D4                    NA         NA  R5-D4 is NA in dog years.
##  9 Biggs Darklighter        24        168  Biggs Darklighter is 168 in dog year~
## 10 Obi-Wan Kenobi           57        399  Obi-Wan Kenobi is 399 in dog years.
## # ... with 77 more rows
```

```
starwars %>%
  select(name, height) %>%
  filter(name %in% c("Luke Skywalker", "Anakin Skywalker")) %>%
  mutate(tall1 = height > 180) %>%
  mutate(tall2 = ifelse(height > 180, "Tall", "Short"))
```

```
## # A tibble: 2 x 4
##   name             height tall1 tall2
##   <chr>             <int> <lgl> <chr>
## 1 Luke Skywalker      172 FALSE Short
## 2 Anakin Skywalker    188 TRUE  Tall
```

Lastly, combining mutate with the new across feature in dplyr 1.0.0+ allows you to easily work on a subset of variables. For example:

```
starwars %>%
  select(name:eye_color) %>%
  mutate(across(where(is.character), toupper)) %>%
  head(5)
```

```
## # A tibble: 5 x 6
##   name           height  mass hair_color skin_color  eye_color
##   <chr>           <int> <dbl> <chr>      <chr>       <chr>
## 1 LUKE SKYWALKER    172    77 BLOND      FAIR        BLUE
## 2 C-3PO             167    75 <NA>       GOLD        YELLOW
## 3 R2-D2              96    32 <NA>       WHITE, BLUE RED
## 4 DARTH VADER       202   136 NONE       WHITE       YELLOW
## 5 LEIA ORGANA       150    49 BROWN      LIGHT       BROWN
```

5. summarise

Collapse multiple rows into a single summary value.

```
starwars %>%
  group_by(species, gender) %>%
  summarise(mean_height = mean(height, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'species'. You can override using the `.groups` argument.
```

```
## # A tibble: 42 x 3
## # Groups:   species [38]
##    species   gender    mean_height
##    <chr>     <chr>           <dbl>
##  1 Aleena    masculine          79
##  2 Besalisk  masculine         198
##  3 Cerean    masculine         198
##  4 Chagrian  masculine         196
##  5 Clawdite  feminine          168
##  6 Droid     feminine           96
##  7 Droid     masculine         140
##  8 Dug       masculine         112
##  9 Ewok      masculine          88
## 10 Geonosian masculine         183
## # ... with 32 more rows
```

Note that including "na.rm = TRUE" (or, its alias "na.rm = T") is usually a good idea with summarise functions. Otherwise, any missing value will propogate to the summarised value too.

```
starwars %>%
  summarise(mean_height = mean(height))
```

```
## # A tibble: 1 x 1
##   mean_height
##         <dbl>
## 1          NA
```

```
starwars %>%
  summarise(mean_height = mean(height, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   mean_height
##         <dbl>
## 1        174.
```

```
starwars %>%
  group_by(species) %>%
  summarise(across(where(is.numeric), mean, na.rm=T)) %>%
  head(5)
```

```
## # A tibble: 5 x 4
##   species  height  mass birth_year
##   <chr>     <dbl> <dbl>      <dbl>
## 1 Aleena       79    15        NaN
## 2 Besalisk    198   102        NaN
## 3 Cerean      198    82         92
## 4 Chagrian    196   NaN        NaN
## 5 Clawdite    168    55        NaN
```

join

```
left_join(flights, planes) %>%
  select(year, month, day, dep_time, arr_time, carrier, flight, tailnum, type, model)
```

```
## Joining, by = c("year", "tailnum")
```

```
## # A tibble: 336,776 x 10
##     year month   day dep_time arr_time carrier flight tailnum type  model
##    <int> <int> <int>    <int>    <int> <chr>    <int> <chr>   <chr> <chr>
##  1  2013     1     1      517      830 UA        1545 N14228  <NA>  <NA>
##  2  2013     1     1      533      850 UA        1714 N24211  <NA>  <NA>
##  3  2013     1     1      542      923 AA        1141 N619AA  <NA>  <NA>
##  4  2013     1     1      544     1004 B6         725 N804JB  <NA>  <NA>
##  5  2013     1     1      554      812 DL         461 N668DN  <NA>  <NA>
##  6  2013     1     1      554      740 UA        1696 N39463  <NA>  <NA>
##  7  2013     1     1      555      913 B6         507 N516JB  <NA>  <NA>
##  8  2013     1     1      557      709 EV        5708 N829AS  <NA>  <NA>
##  9  2013     1     1      557      838 B6          79 N593JB  <NA>  <NA>
## 10  2013     1     1      558      753 AA         301 N3ALAA  <NA>  <NA>
## # ... with 336,766 more rows
```