```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## 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(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.2
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(tables)
```

```
## Warning: package 'tables' was built under R version 3.5.3
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 3.5.2
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
#install.packages("nycflights13")
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 3.5.3
```

```
??nycflights13
```

```
## starting httpd help server ...
```

```
##  done
```

```
german.data <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/
german.data")
```

train test split

```
train <- sample_frac(german.data, 0.75)
test <- anti_join(german.data, train)
```

```
## Joining, by = c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8", "V9", "V10", "V11", "V12", "V1
3", "V14", "V15", "V16", "V17", "V18", "V19", "V20", "V21")
```

select, filter

```
select(storms, year, month, contains("1975"))
```

```
## # A tibble: 10,010 x 2
##     year month
##    <dbl> <dbl>
##  1  1975     6
##  2  1975     6
##  3  1975     6
##  4  1975     6
##  5  1975     6
##  6  1975     6
##  7  1975     6
##  8  1975     6
##  9  1975     6
## 10  1975     6
## # ... with 10,000 more rows
```

```
filter(storms, year>1980)
```

```
## # A tibble: 9,303 x 13
##    name   year month   day  hour   lat  long status category  wind pressure
##    <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>  <ord>    <int>    <int>
##  1 Emily  1981     9     1    12  30.4 -67.3 tropi~ 0           40      996
##  2 Emily  1981     9     1    18  31.3 -66.6 tropi~ 0           45      994
##  3 Emily  1981     9     2     0  31.9 -65.9 tropi~ 0           50      992
##  4 Emily  1981     9     2     6  32.6 -65.1 tropi~ 0           50      990
##  5 Emily  1981     9     2    12  33.3 -64.4 tropi~ 0           50      988
##  6 Emily  1981     9     2    18  34.1 -64.1 tropi~ 0           55      986
##  7 Emily  1981     9     3     0  35   -64   tropi~ 0           55      984
##  8 Emily  1981     9     3     6  36   -65   tropi~ 0           60      982
##  9 Emily  1981     9     3    12  35   -65.8 tropi~ 0           60      980
## 10 Emily  1981     9     3    18  34.2 -65   tropi~ 0           60      978
## # ... with 9,293 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

```
dim(filter(storms, name %in% "Emily"))
```

```
## [1] 207  13
```

Renaming variable

```
test <- test %>% data.table::setnames("sex","gender", skip_absent = TRUE)
test <- test %>% data.table::setnames("gender","sex", skip_absent = TRUE)
colnames(test)
```

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11"
## [12] "V12" "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21"
```

mutate

```
mutate(storms, new = wind/pressure)
```

```
## # A tibble: 10,010 x 14
##    name   year month   day  hour   lat  long status category  wind pressure
##    <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>  <ord>    <int>    <int>
##  1 Amy    1975     6    27     0  27.5 -79   tropi~ -1          25     1013
##  2 Amy    1975     6    27     6  28.5 -79   tropi~ -1          25     1013
##  3 Amy    1975     6    27    12  29.5 -79   tropi~ -1          25     1013
##  4 Amy    1975     6    27    18  30.5 -79   tropi~ -1          25     1013
##  5 Amy    1975     6    28     0  31.5 -78.8 tropi~ -1          25     1012
##  6 Amy    1975     6    28     6  32.4 -78.7 tropi~ -1          25     1012
##  7 Amy    1975     6    28    12  33.3 -78   tropi~ -1          25     1011
##  8 Amy    1975     6    28    18  34   -77   tropi~ -1          30     1006
##  9 Amy    1975     6    29     0  34.4 -75.8 tropi~ 0           35     1004
## 10 Amy    1975     6    29     6  34   -74.8 tropi~ 0           40     1002
## # ... with 10,000 more rows, and 3 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>, new <dbl>
```

summarize

```
summarise(storms, median = median(wind), mean = mean(pressure))
```

```
## # A tibble: 1 x 2
##   median  mean
##    <dbl> <dbl>
## 1     45  992.
```

sort ascending

```
arrange(storms, pressure)
```

```
## # A tibble: 10,010 x 13
##    name   year month   day  hour   lat  long status category  wind pressure
##    <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>  <ord>    <int>    <int>
##  1 Wilma  2005    10    19    12  17.3 -82.8 hurri~ 5          160      882
##  2 Gilb~  1988     9    14     0  19.7 -83.8 hurri~ 5          160      888
##  3 Gilb~  1988     9    14     6  19.9 -85.3 hurri~ 5          155      889
##  4 Gilb~  1988     9    14    12  20.4 -86.5 hurri~ 5          145      892
##  5 Wilma  2005    10    19     6  17   -82.2 hurri~ 5          150      892
##  6 Wilma  2005    10    19    18  17.4 -83.4 hurri~ 5          140      892
##  7 Wilma  2005    10    20     0  17.9 -84   hurri~ 4          135      892
##  8 Rita   2005     9    22     3  24.7 -87.3 hurri~ 5          155      895
##  9 Rita   2005     9    22     0  24.5 -86.9 hurri~ 5          150      897
## 10 Rita   2005     9    22     6  24.8 -87.6 hurri~ 5          155      897
## # ... with 10,000 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

sort descending

```
arrange(storms, desc(pressure))
```

```
## # A tibble: 10,010 x 13
##    name   year month   day  hour   lat  long status category  wind pressure
##    <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>  <ord>    <int>    <int>
##  1 AL07~  2003     7    26    12  32.3 -82   tropi~ -1          20     1022
##  2 AL07~  2003     7    26    18  32.8 -82.6 tropi~ -1          15     1022
##  3 AL07~  2003     7    27     0  33   -83   tropi~ -1          15     1022
##  4 Emily  1993     8    22    18  19.9 -52.6 tropi~ -1          30     1020
##  5 Emily  1993     8    23     0  20.5 -53.6 tropi~ -1          30     1020
##  6 Emily  1993     8    23     6  21.3 -54.8 tropi~ -1          30     1020
##  7 Emily  1993     8    23    12  22.3 -56   tropi~ -1          30     1020
##  8 Emily  1993     8    23    18  23.2 -57.1 tropi~ -1          30     1020
##  9 Emily  1993     8    24     0  24.3 -57.8 tropi~ -1          30     1020
## 10 Emily  1993     8    24     6  25.4 -58.6 tropi~ -1          30     1020
## # ... with 10,000 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

The pipe operator

```
storms%>%
  filter(wind>25) %>%
  select(month:wind) %>%
  arrange(wind)
```

```
## # A tibble: 8,974 x 8
##    month   day  hour   lat  long status              category  wind
##    <dbl> <int> <dbl> <dbl> <dbl> <chr>               <ord>    <int>
##  1     6    28    18  34   -77   tropical depression -1          30
##  2     8    29     0  23   -91.9 tropical depression -1          30
##  3     9     1     0  25.1 -98.3 tropical depression -1          30
##  4     8     6    18  26   -73.4 tropical depression -1          30
##  5     9    27     6  24.9 -58.1 tropical depression -1          30
##  6    10     4     6  36   -43   tropical depression -1          30
##  7    10     4    12  36.5 -41.4 tropical depression -1          30
##  8    10     4    18  37.4 -38.7 tropical depression -1          30
##  9     8    30     0  26.9 -89.4 tropical depression -1          30
## 10     9     7     0  34.4 -75.8 tropical depression -1          30
## # ... with 8,964 more rows
```

Group by

```
storms %>%
  group_by(status) %>%
  select(status, everything(), -name) %>%
  arrange(year)
```

```
## # A tibble: 10,010 x 12
## # Groups:   status [3]
##    status  year month   day  hour   lat  long category  wind pressure
##    <chr>  <dbl> <dbl> <int> <dbl> <dbl> <dbl> <ord>    <int>    <int>
##  1 tropi~  1975     6    27     0  27.5 -79   -1          25     1013
##  2 tropi~  1975     6    27     6  28.5 -79   -1          25     1013
##  3 tropi~  1975     6    27    12  29.5 -79   -1          25     1013
##  4 tropi~  1975     6    27    18  30.5 -79   -1          25     1013
##  5 tropi~  1975     6    28     0  31.5 -78.8 -1          25     1012
##  6 tropi~  1975     6    28     6  32.4 -78.7 -1          25     1012
##  7 tropi~  1975     6    28    12  33.3 -78   -1          25     1011
##  8 tropi~  1975     6    28    18  34   -77   -1          30     1006
##  9 tropi~  1975     6    29     0  34.4 -75.8 0           35     1004
## 10 tropi~  1975     6    29     6  34   -74.8 0           40     1002
## # ... with 10,000 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

```
storms %>%
  group_by(status) %>%
  summarise(mean = mean(wind), median=median(lat))
```

```
## # A tibble: 3 x 3
##   status               mean median
##   <chr>               <dbl>  <dbl>
## 1 hurricane            86.0   26.2
## 2 tropical depression  27.3   21.6
## 3 tropical storm       45.8   24.4
```

Ungroup

```
storms %>%
  ungroup()
```

```
## # A tibble: 10,010 x 13
##    name   year month   day  hour   lat  long status category  wind pressure
##    <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>  <ord>    <int>    <int>
##  1 Amy    1975     6    27     0  27.5 -79   tropi~ -1          25     1013
##  2 Amy    1975     6    27     6  28.5 -79   tropi~ -1          25     1013
##  3 Amy    1975     6    27    12  29.5 -79   tropi~ -1          25     1013
##  4 Amy    1975     6    27    18  30.5 -79   tropi~ -1          25     1013
##  5 Amy    1975     6    28     0  31.5 -78.8 tropi~ -1          25     1012
##  6 Amy    1975     6    28     6  32.4 -78.7 tropi~ -1          25     1012
##  7 Amy    1975     6    28    12  33.3 -78   tropi~ -1          25     1011
##  8 Amy    1975     6    28    18  34   -77   tropi~ -1          30     1006
##  9 Amy    1975     6    29     0  34.4 -75.8 tropi~ 0           35     1004
## 10 Amy    1975     6    29     6  34   -74.8 tropi~ 0           40     1002
## # ... with 10,000 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

Join data

```
a <- data.frame("First Name" = c("Anu", "Priya", "Kichu"), "Last Name" = c("Kush", "Singhania", "Ol
iver"), "Salary" = c(56000,23000,89999))
b <- data.frame("Age" = 23:25, "SEX" = c("F", "F", "M"), "Country" = c("US", "India", "Sri Lanka"))
new.data <- bind_cols(a,b)
new.data
```

```
##   First.Name Last.Name Salary Age SEX   Country
## 1        Anu      Kush  56000  23   F        US
## 2      Priya Singhania  23000  24   F     India
## 3      Kichu    Oliver  89999  25   M Sri Lanka
```

```
select(new.data, Country, everything())
```

```
##     Country First.Name Last.Name Salary Age SEX
## 1        US        Anu      Kush  56000  23   F
## 2     India      Priya Singhania  23000  24   F
## 3 Sri Lanka      Kichu    Oliver  89999  25   M
```

```
a <- data.frame("First Name" = c("Anu", "Priya", "Kichu"), "Last Name" = c("Kush", "Singhania", "Ol
iver"), "Salary" = c(56000,23000,89999))
b <- data.frame("Age" = 23:25, "SEX" = c("F", "F", "M"), "Package" = c(45000,90000,23000))
new.data <- bind_rows(a,b)
new.data
```

```
##   First.Name Last.Name Salary Age  SEX Package
## 1        Anu      Kush  56000  NA <NA>      NA
## 2      Priya Singhania  23000  NA <NA>      NA
## 3      Kichu    Oliver  89999  NA <NA>      NA
## 4       <NA>      <NA>     NA  23    F   45000
## 5       <NA>      <NA>     NA  24    F   90000
## 6       <NA>      <NA>     NA  25    M   23000
```

```
select(new.data, everything())
```

```
##   First.Name Last.Name Salary Age  SEX Package
## 1        Anu      Kush  56000  NA <NA>      NA
## 2      Priya Singhania  23000  NA <NA>      NA
## 3      Kichu    Oliver  89999  NA <NA>      NA
## 4       <NA>      <NA>     NA  23    F   45000
## 5       <NA>      <NA>     NA  24    F   90000
## 6       <NA>      <NA>     NA  25    M   23000
```

Join

```
left_join(a, b, by = c("Salary"="Package"))
```

```
##   First.Name Last.Name Salary Age  SEX
## 1        Anu      Kush  56000  NA <NA>
## 2      Priya Singhania  23000  25    M
## 3      Kichu    Oliver  89999  NA <NA>
```