Data Exploration/Visualization in R

Fabienne ishimwe
11/12/2017

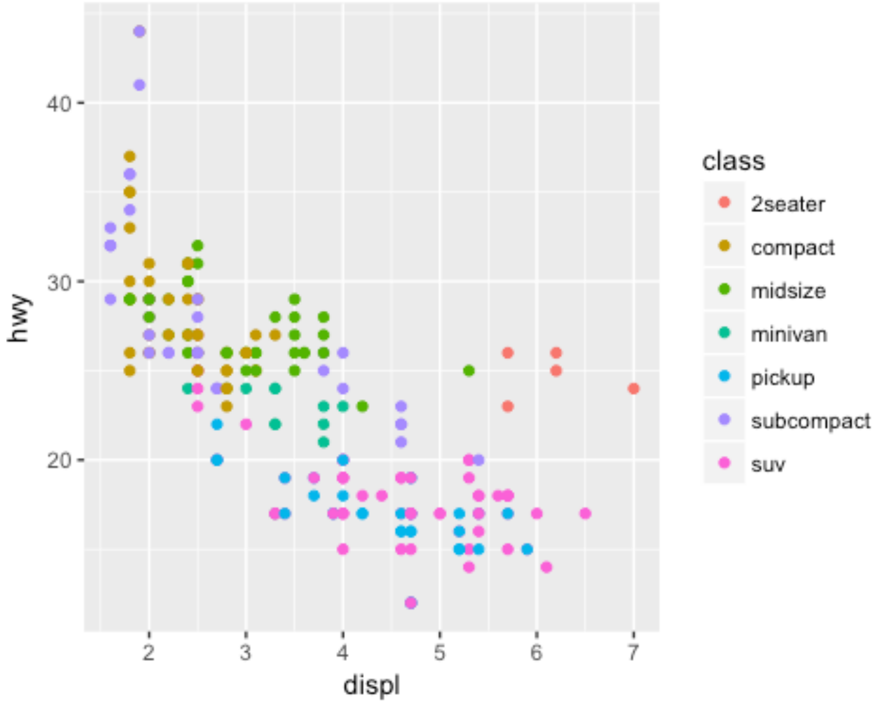
Loading libraries

```
library('tidyverse')
```

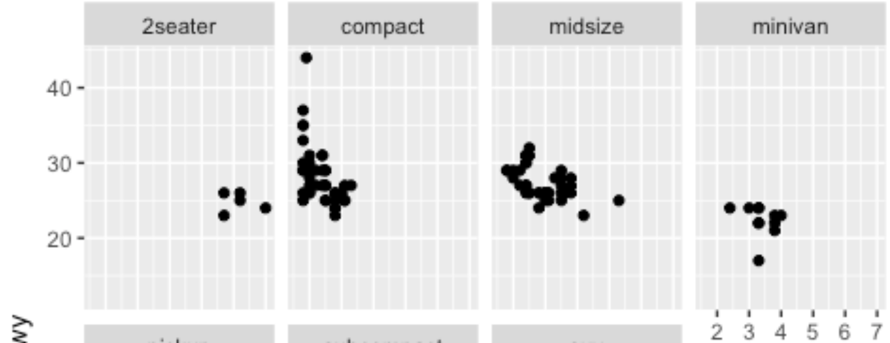
```
## ── Attaching packages ──────── tidyverse 1.2.0 ──
## ✔ ggplot2 2.2.1     ✔ purrr   0.2.4
## ✔ tibble  1.3.4     ✔ dplyr   0.7.4
## ✔ tidyr   0.7.2     ✔ stringr 1.2.0
## ✔ readr   1.1.1     ✔ forcats 0.2.0
## ── Conflicts ───────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

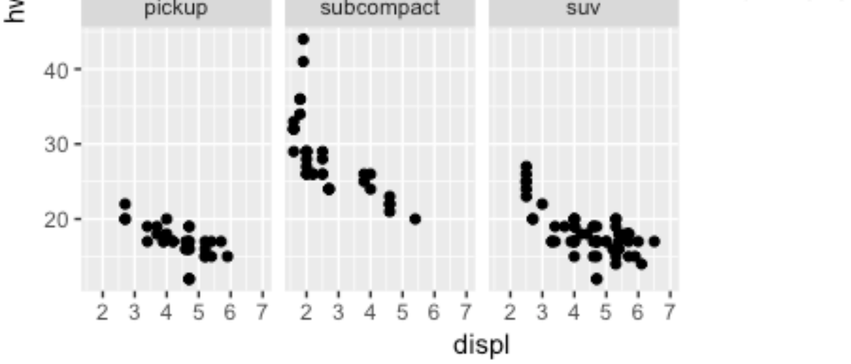
Data Visualization

```
# ggplot simple template
#==============================================
#ggplot(data = <DATA>) +
 #<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
#==============================================

# Visualizing with ggplot: plot with no facets
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = class))
```

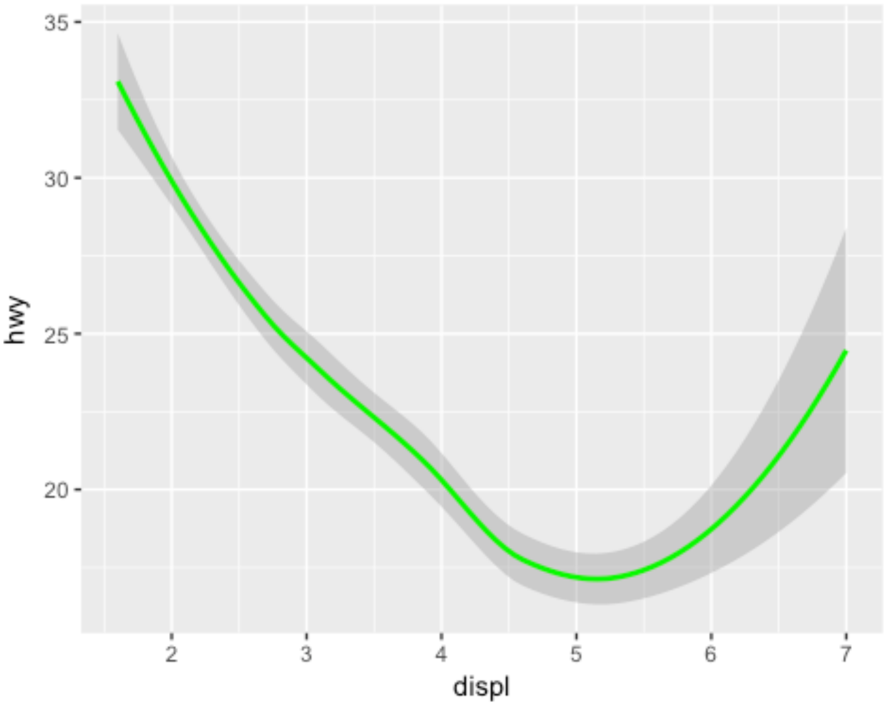


```
## Visualizing with ggplot: plot wit subplots/facets
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))+
facet_wrap(~class, nrow=2)
```

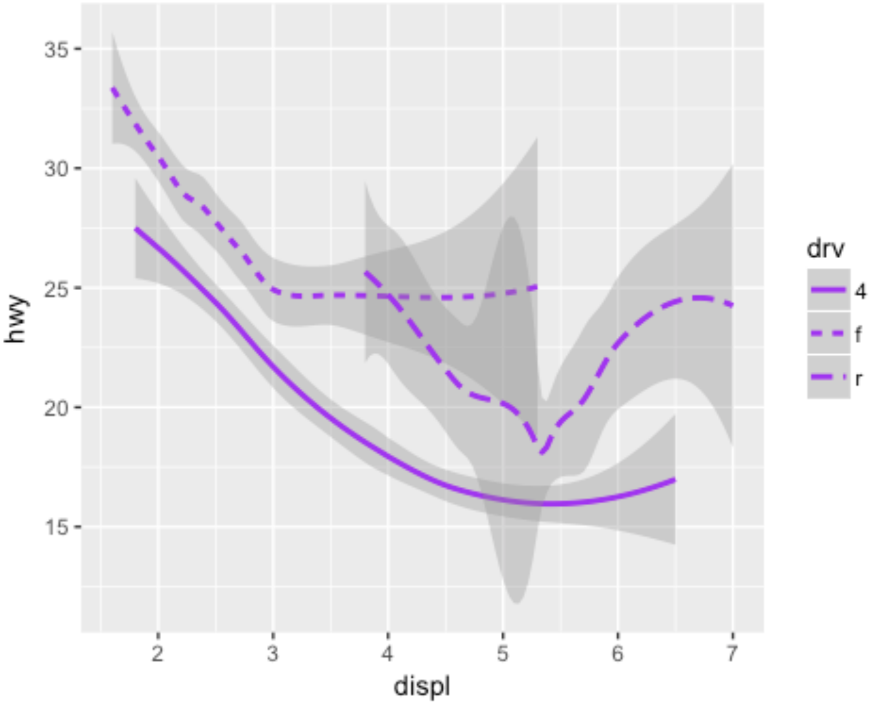




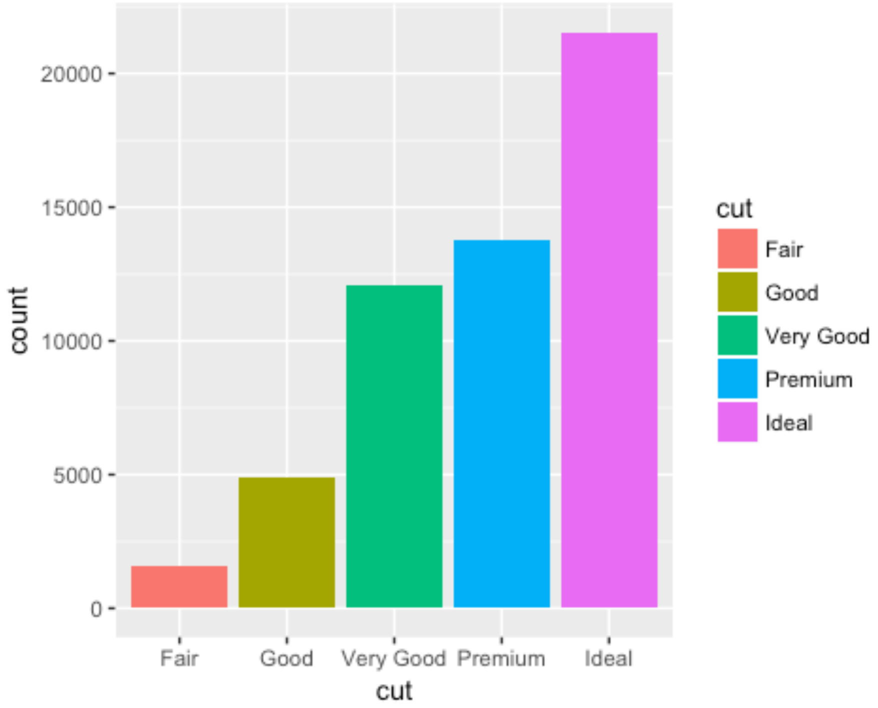
```
# Making line graph
ggplot(data = mpg) +
 geom_smooth(mapping = aes(x = displ, y = hwy), color='green')
## `geom_smooth()` using method = 'loess'
```



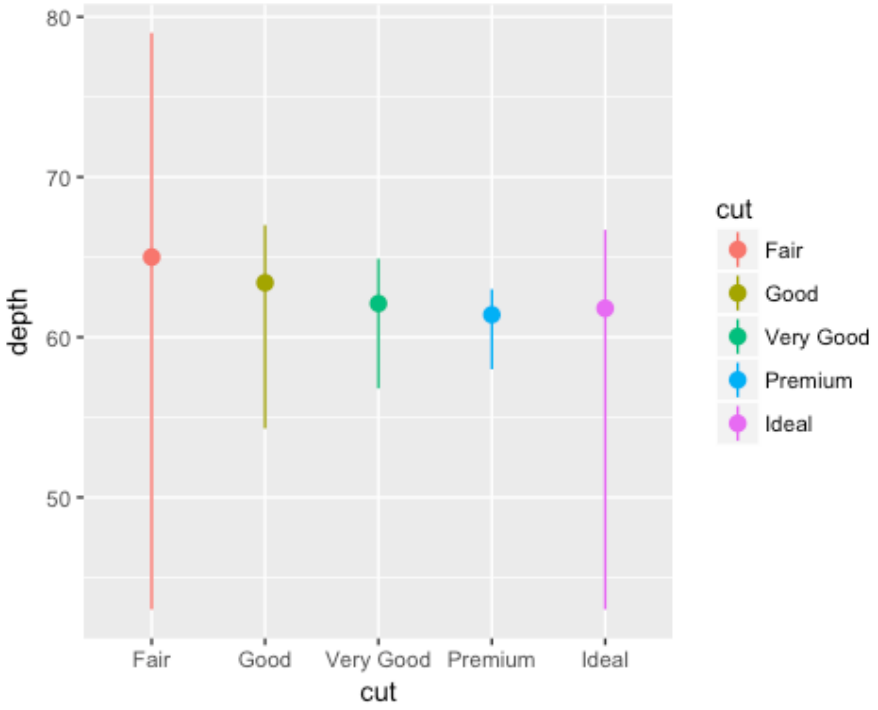
```
# Classify  the data by drv and visualizing it with different lines
ggplot(data = mpg) +
 geom_smooth(mapping = aes(x = displ, y = hwy, linetype=drv), color='purple')
## `geom_smooth()` using method = 'loess'
```



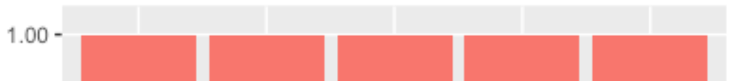
```
# bar plot

ggplot(data = diamonds) +
 geom_bar(mapping = aes(x = cut, fill = cut))
```

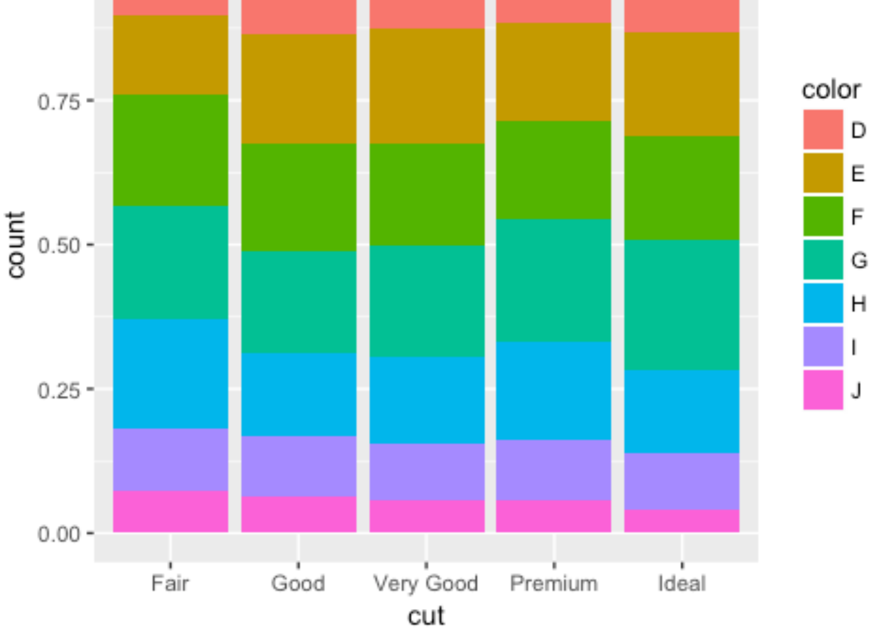


```
# summarizing the depth value for each cut types
ggplot(data = diamonds) +
 stat_summary(
   mapping = aes(x = cut, y = depth, color=cut),
   fun.ymin = min,
   fun.ymax = max,
   fun.y = median
 )
```

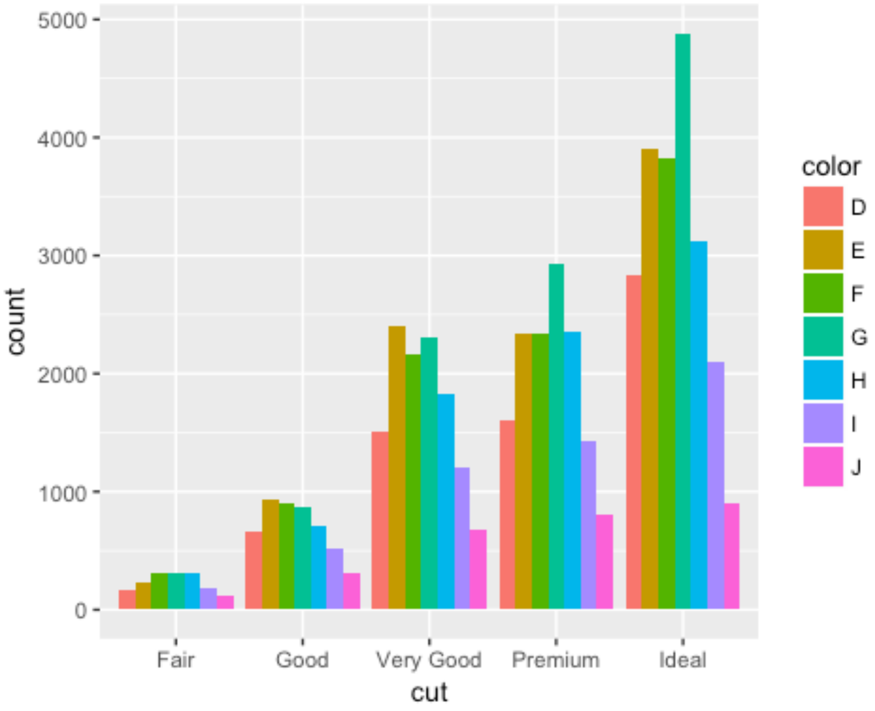


```
# Stacked bar plot by "clarity"
ggplot(data = diamonds) +
 geom_bar(mapping = aes(x = cut, fill =color),position = "fill") # position makes the size of the bars the same so one can compare
```





```
# using positon dodge for side by side bars instead
ggplot(data = diamonds) +
 geom_bar(mapping = aes(x = cut, fill = color), position = "dodge")
```



Data Transformation

```
#install.packages('nycflights13')
library(nycflights13)
# using filter to subset a dataframe
filter(flights, month == 1, day == 1)
## # A tibble: 842 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 832 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
# using near to
near(sqrt(2) ^ 2,  2)
## [1] TRUE
# logical 1
filter(flights, month == 11 | month == 12)
## # A tibble: 55,403 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013    11     1        5           2359         6      352
##  2  2013    11     1       35           2250       105      123
##  3  2013    11     1      455            500        -5      641
##  4  2013    11     1      539            545        -6      856
##  5  2013    11     1      542            545        -3      831
##  6  2013    11     1      549            600       -11      912
##  7  2013    11     1      550            600       -10      705
##  8  2013    11     1      554            600        -6      659
##  9  2013    11     1      554            600        -6      826
## 10  2013    11     1      554            600        -6      749
## # ... with 55,393 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
# logical 1 same as  logical 2

# logical 2
filter(flights, month %in% c(11, 12))
## # A tibble: 55,403 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013    11     1        5           2359         6      352
##  2  2013    11     1       35           2250       105      123
##  3  2013    11     1      455            500        -5      641
##  4  2013    11     1      539            545        -6      856
##  5  2013    11     1      542            545        -3      831
##  6  2013    11     1      549            600       -11      912
##  7  2013    11     1      550            600       -10      705
##  8  2013    11     1      554            600        -6      659
##  9  2013    11     1      554            600        -6      826
## 10  2013    11     1      554            600        -6      749
## # ... with 55,393 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

Missing Values

```
df=tibble(x = c(1, NA, 3,4))

filter(df,is.na(x) |x>1)
## # A tibble: 3 x 1
##       x
##   <dbl>
## 1    NA
## 2     3
## 3     4
# Arranging values in a dataframe
arrange(flights, desc(year),month, day)
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
# Selecting columns by name
```

```
select(flights,month, day)
## # A tibble: 336,776 x 2
##    month   day
##    <int> <int>
##  1     1     1
##  2     1     1
##  3     1     1
##  4     1     1
##  5     1     1
##  6     1     1
##  7     1     1
##  8     1     1
##  9     1     1
## 10     1     1
## # ... with 336,766 more rows
# Selecting all columns except those from year to day (inclusive)
select(flights, -(year:day))
## # A tibble: 336,776 x 16
##    dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
##       <int>          <int>     <dbl>    <int>          <int>     <dbl>
##  1      517            515         2      830            819        11
##  2      533            529         4      850            830        20
##  3      542            540         2      923            850        33
##  4      544            545        -1     1004           1022       -18
##  5      554            600        -6      812            837       -25
##  6      554            558        -4      740            728        12
##  7      555            600        -5      913            854        19
##  8      557            600        -3      709            723       -14
##  9      557            600        -3      838            846        -8
## 10      558            600        -2      753            745         8
## # ... with 336,766 more rows, and 10 more variables: carrier <chr>,
## #   flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
#renaming columns in a select
rename(flights, tail_num = tailnum)
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tail_num <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
# Adding new columns with mutate
flights_small <- select(flights,
 year:day,
 ends_with("delay"),
 distance,
 air_time
)
mutate(flights_small,
 gain = arr_delay - dep_delay,
 speed = distance / air_time * 30
)
## # A tibble: 336,776 x 9
##     year month   day dep_delay arr_delay distance air_time  gain    speed
##    <int> <int> <int>     <dbl>     <dbl>    <dbl>    <dbl> <dbl>    <dbl>
##  1  2013     1     1         2        11     1400      227     9 185.0220
##  2  2013     1     1         4        20     1416      227    16 187.1366
##  3  2013     1     1         2        33     1089      160    31 204.1875
##  4  2013     1     1        -1       -18     1576      183   -17 258.3607
##  5  2013     1     1        -6       -25      762      116   -19 197.0690
##  6  2013     1     1        -4        12      719      150    16 143.8000
##  7  2013     1     1        -5        19     1065      158    24 202.2152
##  8  2013     1     1        -3       -14      229       53   -11 129.6226
##  9  2013     1     1        -3        -8      944      140    -5 202.2857
## 10  2013     1     1        -2         8      733      138    10 159.3478
## 10  2013     1     1        -2         8       755       158     10 139.5478
## # ... with 336,766 more rows
```