Untitled

Zichun Liu

9/25/2017

```
microbenchmark(
  filter(flights, month == 1, day == 1),
  subset(flights, subset = month == 1 & day ==1)
)
```

```
## Unit: milliseconds
##                                           expr       min        lq
##            filter(flights, month == 1, day == 1)  6.139183  7.957064
##  subset(flights, subset = month == 1 & day == 1) 13.253462 17.208868
##      mean    median       uq       max neval cld
##  15.32484  9.328436 10.76085 571.35539   100   a
##  21.67645 20.991920 23.55918  78.21059   100   a
```

```
#select
flights %>%
  select(month, day, carrier, distance)
```

```
## # A tibble: 336,776 × 4
##    month   day carrier distance
##    <int> <int>   <chr>    <dbl>
## 1      1     1      UA     1400
## 2      1     1      UA     1416
## 3      1     1      AA     1089
## 4      1     1      B6     1576
## 5      1     1      DL      762
## 6      1     1      UA      719
## 7      1     1      B6     1065
## 8      1     1      EV      229
## 9      1     1      B6      944
## 10     1     1      AA      733
## # ... with 336,766 more rows
```

```
#filter
flights %>%
  select(month, day, carrier, distance) %>%
  filter(month == 1, day == 1)
```

```
## # A tibble: 842 × 4
##    month   day carrier distance
##    <int> <int>   <chr>    <dbl>
## 1      1     1      UA     1400
## 2      1     1      UA     1416
## 3      1     1      AA     1089
## 4      1     1      B6     1576
## 5      1     1      DL      762
## 6      1     1      UA      719
## 7      1     1      B6     1065
## 8      1     1      EV      229
## 9      1     1      B6      944
```

```
## 10     1     1      AA      733
## # ... with 832 more rows
```

```
# group_by and summarise
flights %>%
  select(month, day, carrier, distance) %>%
  filter(month == 1, day == 1) %>%
  group_by(carrier) %>%
  summarise(mean(distance),n())
```

```
## # A tibble: 14 × 3
##    carrier `mean(distance)` `n()`
##      <chr>            <dbl> <int>
## 1       9E         520.3571    28
## 2       AA        1337.7128    94
## 3       AS        2402.0000     2
## 4       B6        1106.2025   163
## 5       DL        1222.0357   112
## 6       EV         491.4569   116
## 7       F9        1620.0000     2
## 8       FL         686.6000    10
## 9       HA        4983.0000     1
## 10      MQ         577.0000    78
## 11      UA        1496.4909   165
## 12      US         833.1562    32
## 13      VX        2502.3333    12
## 14      WN         895.7037    27
```

```
#arrange()
flights %>%
  select(month, day, carrier, distance) %>%
  filter(month == 1, day == 1) %>%
  group_by(carrier) %>%
  summarise(meandistance = mean(distance),n()) %>%
  arrange(meandistance)
```

```
## # A tibble: 14 × 3
##    carrier meandistance `n()`
##      <chr>        <dbl> <int>
## 1       EV     491.4569   116
## 2       9E     520.3571    28
## 3       MQ     577.0000    78
## 4       FL     686.6000    10
## 5       US     833.1562    32
## 6       WN     895.7037    27
## 7       B6    1106.2025   163
## 8       DL    1222.0357   112
## 9       AA    1337.7128    94
## 10      UA    1496.4909   165
## 11      F9    1620.0000     2
## 12      AS    2402.0000     2
## 13      VX    2502.3333    12
## 14      HA    4983.0000     1
```

```
# mutate
flights %>%
```

```
  select(month, day, carrier, distance) %>%
  filter(month == 1, day == 1) %>%
  group_by(carrier) %>%
  summarise(meandistance = mean(distance), N = n()) %>%
  arrange(meandistance) %>%
  mutate(totaldistance = meandistance*N)
```

```
## # A tibble: 14 × 4
##    carrier meandistance     N totaldistance
##      <chr>        <dbl> <int>         <dbl>
## 1       EV     491.4569   116         57009
## 2       9E     520.3571    28         14570
## 3       MQ     577.0000    78         45006
## 4       FL     686.6000    10          6866
## 5       US     833.1562    32         26661
## 6       WN     895.7037    27         24184
## 7       B6    1106.2025   163        180311
## 8       DL    1222.0357   112        136868
## 9       AA    1337.7128    94        125745
## 10      UA    1496.4909   165        246921
## 11      F9    1620.0000     2          3240
## 12      AS    2402.0000     2          4804
## 13      VX    2502.3333    12         30028
## 14      HA    4983.0000     1          4983
```

```
# join
flight_1 = flights %>%
  select(month, day, carrier, distance) %>%
  filter(month == 1, day == 1) %>%
  group_by(carrier) %>%
  summarise(meandistance = mean(distance), N = n()) %>%
  arrange(meandistance) %>%
  mutate(totaldistance = meandistance*N, rank = 1:length(meandistance))


flight_2 = flights %>%
  select(month, day, carrier,air_time) %>%
  filter(month == 1, day == 1) %>%
  group_by(carrier) %>%
  summarise(meanairtime = mean(air_time)) %>%
  na.omit()

inner_join(flight_1, flight_2, "carrier")
```

```
## # A tibble: 8 × 6
##   carrier meandistance     N totaldistance  rank meanairtime
##     <chr>        <dbl> <int>         <dbl> <int>       <dbl>
## 1      FL     686.6000    10          6866     4    120.5000
## 2      US     833.1562    32         26661     5    139.6250
## 3      WN     895.7037    27         24184     6    162.8148
## 4      DL    1222.0357   112        136868     8    184.7768
## 5      F9    1620.0000     2          3240    11    249.5000
## 6      AS    2402.0000     2          4804    12    337.0000
## 7      VX    2502.3333    12         30028    13    353.6667
## 8      HA    4983.0000     1          4983    14    659.0000
```

```
left_join(flight_1, flight_2, "carrier")
```

```
## # A tibble: 14 × 6
##    carrier meandistance     N totaldistance  rank meanairtime
##      <chr>        <dbl> <int>         <dbl> <int>       <dbl>
## 1       EV     491.4569   116         57009     1          NA
## 2       9E     520.3571    28         14570     2          NA
## 3       MQ     577.0000    78         45006     3          NA
## 4       FL     686.6000    10          6866     4    120.5000
## 5       US     833.1562    32         26661     5    139.6250
## 6       WN     895.7037    27         24184     6    162.8148
## 7       B6    1106.2025   163        180311     7          NA
## 8       DL    1222.0357   112        136868     8    184.7768
## 9       AA    1337.7128    94        125745     9          NA
## 10      UA    1496.4909   165        246921    10          NA
## 11      F9    1620.0000     2          3240    11    249.5000
## 12      AS    2402.0000     2          4804    12    337.0000
## 13      VX    2502.3333    12         30028    13    353.6667
## 14      HA    4983.0000     1          4983    14    659.0000
```

```
semi_join(flight_1, flight_2, "carrier")
```

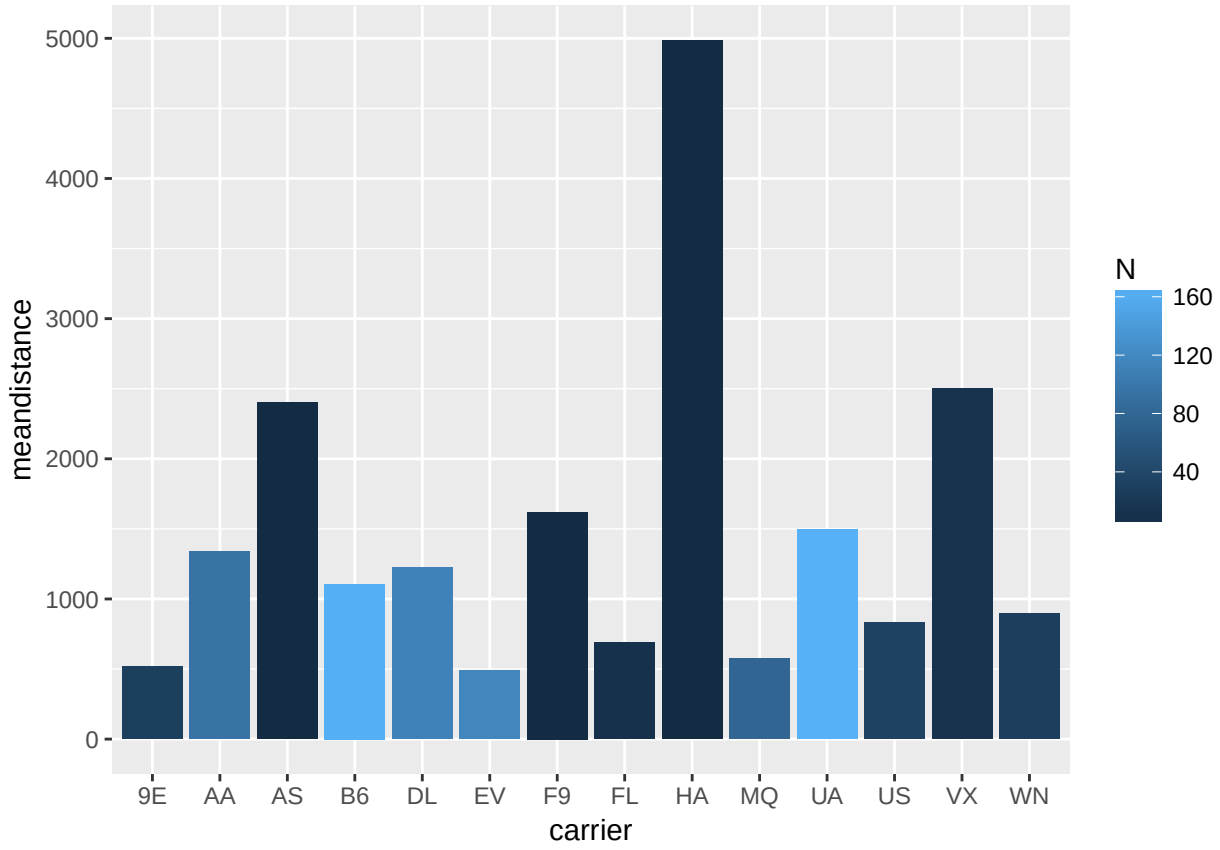
```
## # A tibble: 8 × 5
##   carrier meandistance     N totaldistance  rank
##     <chr>        <dbl> <int>         <dbl> <int>
## 1      AS    2402.0000     2          4804    12
## 2      DL    1222.0357   112        136868     8
## 3      F9    1620.0000     2          3240    11
## 4      FL     686.6000    10          6866     4
## 5      HA    4983.0000     1          4983    14
## 6      US     833.1562    32         26661     5
## 7      VX    2502.3333    12         30028    13
## 8      WN     895.7037    27         24184     6
```

```
anti_join(flight_1, flight_2, "carrier")
```

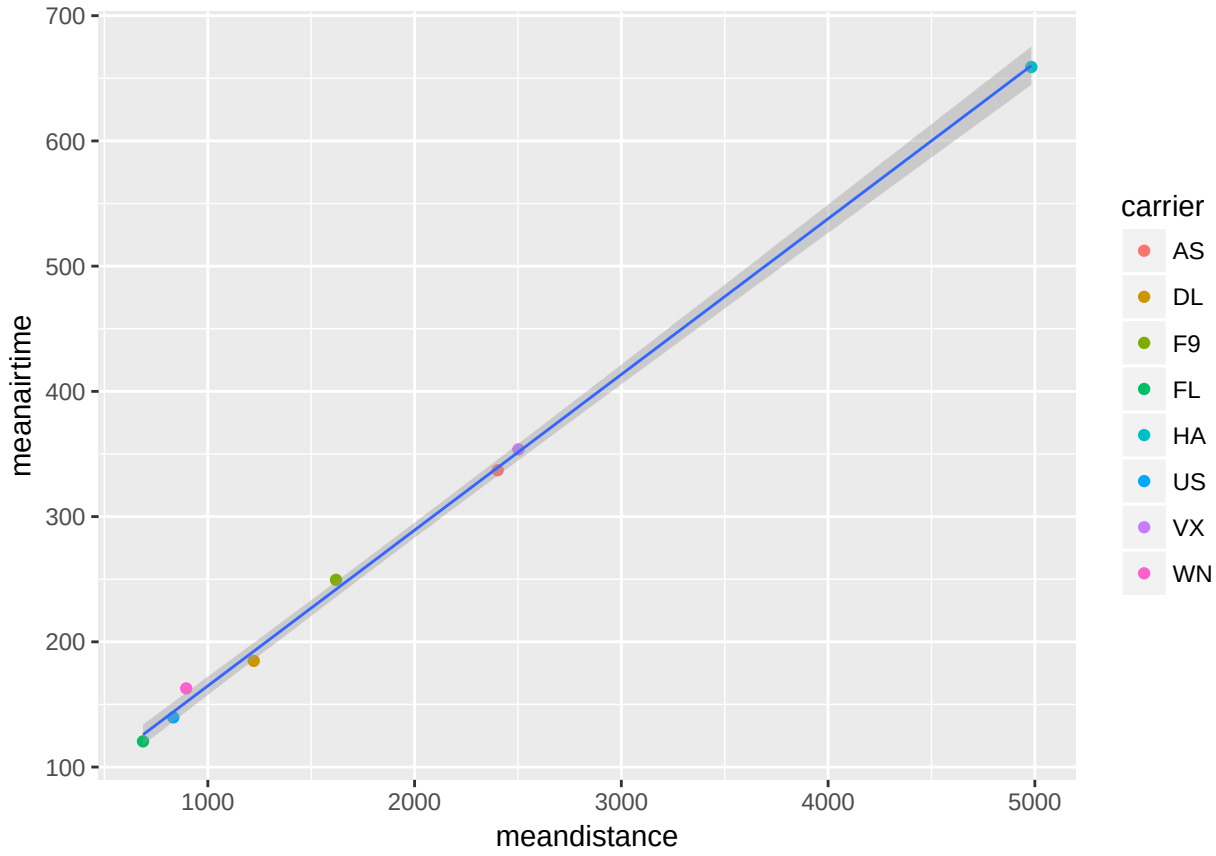
```
## # A tibble: 6 × 5
##   carrier meandistance     N totaldistance  rank
##     <chr>        <dbl> <int>         <dbl> <int>
## 1      UA    1496.4909   165        246921    10
## 2      AA    1337.7128    94        125745     9
## 3      B6    1106.2025   163        180311     7
## 4      MQ     577.0000    78         45006     3
## 5      9E     520.3571    28         14570     2
## 6      EV     491.4569   116         57009     1
```

```
library(ggplot2)
flights %>%
  select(month, day, carrier, distance) %>%
  filter(month == 1, day == 1) %>%
  group_by(carrier) %>%
  summarise(meandistance = mean(distance), N = n()) %>%
  arrange(meandistance) %>%
  mutate(totaldistance = meandistance*N, rank = 1:length(meandistance)) %>%
  ggplot() +
```

```
  geom_bar(aes(x = carrier, y = meandistance, fill = N), stat = "identity")
```




```
inner_join(flight_1, flight_2, "carrier") %>%
  ggplot() +
  geom_point(aes(x = meandistance, y = meanairtime, col = carrier)) +
  geom_smooth(aes(x = meandistance, y = meanairtime), method = "lm", lwd = 0.5)
```


700
600
500
400
300
200
100
meanairtime
1000
2000
3000
4000
5000
meandistance
carrier
AS
DL
F9
FL
HA
US
VX
WN