Dplyr vs Data.Table

Statistics 506, Fall 2017

Instructions

Consider the 2014 flights data used for the `data.table` examples. For each code snippet first provide a single-sentence summary of the value(s) being computed. Make your answer as concise and descriptive as possible.

Questions

1. Consider the following `data.table` expression:

```
tab1 =
  nyc14[ , .(n = .N), by=.(origin, dest)] %>%
    .[, .(origin, n = n, pct = n / sum(n)), by=dest] %>%
    .[pct > .75] %>%
    .[order(-pct, dest)]
```

   a. Provide a one-sentence summary of what is being computed.

   b. Provide a translation using `dplyr` syntax.

2. Consider the `dplyr` code snippet below.

```
tab2 =
  nyc14 %>%
    group_by(origin, dest, carrier) %>%
    summarize(n = n()) %>%
    filter(n >= 80) %>%
    group_by(origin, carrier) %>%
    summarize(n = n()) %>%
    arrange(origin,-n)
```

   a. Provide a one-sentence summary of what is being computed.

   b. Provide a translation using `data.table` syntax.

3. Consider the R code snippet below.

```
nyc14_df = as.data.frame(nyc14)
tab3 =
 with(
  with(nyc14_df,nyc14_df[grepl('HOU',dest),]),
      {
       keys = paste(carrier,month,sep=':')
       u = unique(keys)
       n = sapply(u, function(key) sum(key==keys))

       tmp = strsplit(u,':')
       carrier = sapply(tmp,function(x) x[1])
       month  = sapply(tmp,function(x) x[2])

       cr  = unique(carrier)
       ind = sapply(cr, function(x) grep(x,carrier))
       data.frame(carrier,month,n)[ind,]
      }
 )
```

a. Provide a one-sentence summary of what is being computed.

b. Provide a translation using `data.table` syntax.

c. Provide a translation using `dplyr` syntax.