
Question1 (5 points)

Consider the following data we used in class

```
> tidyr::who
```

and recall we tidied the data in class use the following,

```
> who_tidy_tb = who %>%
+   gather(
+     new_sp_m014:newrel_f65, key = "tmp",
+     value = "counts", na.rm = TRUE
+          ) %>%
+   mutate(
+     tmp = stringr::str_replace(
+       tmp, "newrel", "new_rel")
+     ) %>%
+   separate(
+     col = tmp, sep = "_",
+     into = c("new", "type", "sexage")
+     ) %>%
+   select(-new, -iso2, -iso3) %>%
+   separate(col = sexage,
+            into = c("gender", "age"),
+            sep = 1)
```

(a) (1 point) Convert `tidyr::who`, which is a tibble, into a data.table, name it `who_dt`.

(b) (1 point) Convert `who_dt`, which is wide, into a long format, name it `who_long_dt`.

(c) (1 point) Create new columns `type`, `gender`, and `age` in `who_long_dt`.

(d) (1 point) Select columns `country`, `year`, `type`, `gender`, `age`, and `counts`, name it `nwho_dt`.

(e) (1 point) Use `rbenchmark::benchmark` to compare the two implements. Set `replications=10`.

Question2 (8 points)

Consider the following five datasets provided by the library `nycflights13`,

```
> airlines; airports; flights; planes; weather
```

You might find the following useful when comes to join different relational data together,

```
> help(inner_join)
> help(left_join)
> help(right_join)
> help(full_join)
```

(a) (2 points) Compute the average delay by destination, then join on the `airports` data so you can show the spatial distribution of delays. Here is an easy way to draw a map of the united states:

```
> library(nycflights13); library(dplyr); library(ggplot2)
>
> airports %>%
+   semi_join(flights, c("faa" = "dest")) %>%
+   ggplot(aes(lon, lat)) +
+     borders("state") +
+   geom_point() +
+   ggplot2::coord_quickmap()
```

You might want to use the `size` or `color` of the points to dispaly the average delay for each airport.

(b) (2 points) Add the location of the origin and destination, that is, `lat` and `lon`, to `flights`.

(c) (2 points) Is there any relationship between the age of a plane and its delays?

(d) (2 points) What happened on June 13, 2013? Display the spatial pattern of delays, and cross-reference with the weather.

Question3 (7 points)

Consider the 2008 flight data we used in class, `2008.csv.bz2`, which is about 100MB and 800MB when uncompressed into csv.

(a) (1 point) Setup a `ffdf` folder and read the data into R using `read.table.ffdf`, name it `flights.2008.ff.data`.

(b) (1 point) Use `fread` to read the file and name it as `flights_2008_DT`, and run the following linear regression model

```
> flights.LM =
+    lm(DepDelay~DayOfWeek+DepTime+CRSDepTime+ArrTime+CRSArrTime
+       +UniqueCarrier, data = flights_2008_DT)
```

(c) (2 points) Perform the usual regression analysis to improve the model. i.e. variable selection and simple transformation.

(d) (1 point) Run the final model use `flights.2008.ff.data` instead of `flights_2008_DT`. Use `rbenchmark::benchmark` to compare the two implements.

(e) (1 point) The whole flight data is very big, 1987-2008 along is about 16G, which means your laptop will not be able to handle it as a whole. It can be download Here. Download at least one more year, i.e. 2007, to refine your regression model.

(f) (1 point) Use `rbenchmark::benchmark` to compare the two implements when data size increases. If you have a really powerful laptop, you might have to download a few more years to see the difference.