Homework 3

Manav Mandhani mm58926

This homework is due on Feb 10, 2015 in class.

Question 1: (4 pts) *The datset* `AirPassengers` *built into R lists total numbers of international airline passengers, 1949 to 1960.*

```
AirPassengers
```

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949 112 118 132 129 121 135 148 148 136 119 104 118
## 1950 115 126 141 135 125 149 170 170 158 133 114 140
## 1951 145 150 178 163 172 178 199 199 184 162 146 166
## 1952 171 180 193 181 183 218 230 242 209 191 172 194
## 1953 196 196 236 235 229 243 264 272 237 211 180 201
## 1954 204 188 235 227 234 264 302 293 259 229 203 229
## 1955 242 233 267 269 270 315 364 347 312 274 237 278
## 1956 284 277 317 313 318 374 413 405 355 306 271 306
## 1957 315 301 356 348 355 422 465 467 404 347 305 336
## 1958 340 318 362 348 363 435 491 505 404 359 310 337
## 1959 360 342 406 396 420 472 548 559 463 407 362 405
## 1960 417 391 419 461 472 535 622 606 508 461 390 432
```

Is the dataset tidy? Explain why or why not.

No, it is not tidy. Variables do not correspond to columns and each row doesn't represent one unit but a bunch of units aggregated by year.

The datset `HairEyeColor` *built into R contains the distribution of hair color, eye color, and sex in 592 statistics students.*

```
HairEyeColor
```

```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

Is the dataset tidy? Explain why or why not.

It is not tidy. The column headers are values, not variable names. Eye color should be variable instead of black, brown etc

Question 2: (2 pts) *The function* `data()` *lists all datasets that are available in R by default. Look through the list and identify a dataset that is tidy. Explain why the dataset is tidy.*

I pick the dataset: …

```
morley
```

```
##     Expt Run Speed
## 001    1   1   850
## 002    1   2   740
## 003    1   3   900
## 004    1   4  1070
## 005    1   5   930
## 006    1   6   850
## 007    1   7   950
## 008    1   8   980
## 009    1   9   980
## 010    1  10   880
## 011    1  11  1000
## 012    1  12   980
## 013    1  13   930
## 014    1  14   650
## 015    1  15   760
## 016    1  16   810
## 017    1  17  1000
## 018    1  18  1000
## 019    1  19   960
## 020    1  20   960
```

```
## 021    2   1   960
## 022    2   2   940
## 023    2   3   960
## 024    2   4   940
## 025    2   5   880
## 026    2   6   800
## 027    2   7   850
## 028    2   8   880
## 029    2   9   900
## 030    2  10   840
## 031    2  11   830
## 032    2  12   790
## 033    2  13   810
## 034    2  14   880
## 035    2  15   880
## 036    2  16   830
## 037    2  17   800
## 038    2  18   790
## 039    2  19   760
## 040    2  20   800
## 041    3   1   880
## 042    3   2   880
## 043    3   3   880
## 044    3   4   860
## 045    3   5   720
## 046    3   6   720
## 047    3   7   620
## 048    3   8   860
## 049    3   9   970
## 050    3  10   950
## 051    3  11   880
## 052    3  12   910
## 053    3  13   850
## 054    3  14   870
## 055    3  15   840
## 056    3  16   840
## 057    3  17   850
## 058    3  18   840
## 059    3  19   840
## 060    3  20   840
## 061    4   1   890
## 062    4   2   810
## 063    4   3   810
## 064    4   4   820
## 065    4   5   800
## 066    4   6   770
## 067    4   7   760
## 068    4   8   740
## 069    4   9   750
## 070    4  10   760
```

```
## 071    4  11   910
## 072    4  12   920
## 073    4  13   890
## 074    4  14   860
## 075    4  15   880
## 076    4  16   720
## 077    4  17   840
## 078    4  18   850
## 079    4  19   850
## 080    4  20   780
## 081    5   1   890
## 082    5   2   840
## 083    5   3   780
## 084    5   4   810
## 085    5   5   760
## 086    5   6   810
## 087    5   7   790
## 088    5   8   810
## 089    5   9   820
## 090    5  10   850
## 091    5  11   870
## 092    5  12   870
## 093    5  13   810
## 094    5  14   740
## 095    5  15   810
## 096    5  16   940
## 097    5  17   950
## 098    5  18   800
## 099    5  19   810
## 100    5  20   870
```

Explanation goes here.

Each run is in a row, column headers represent variables.

Question 3: (2 pts) *The package nycflights13 contains information about all flights departing from one of the NY City airports in 2013. In particular, the data table* `flights` *lists on-time departure and arrival information for 336,776 individual flights:*

```
library(nycflights13)
head(flights)
```

```
## Source: local data frame [6 x 16]
##
##   year month day dep_time dep_delay arr_time arr_delay carrier tailnum
## 1 2013     1   1      517         2      830        11      UA  N14228
## 2 2013     1   1      533         4      850        20      UA  N24211
## 3 2013     1   1      542         2      923        33      AA  N619AA
## 4 2013     1   1      544        -1     1004       -18      B6  N804JB
## 5 2013     1   1      554        -6      812       -25      DL  N668DN
## 6 2013     1   1      554        -4      740        12      UA  N39463
## Variables not shown: flight (int), origin (chr), dest (chr), air_time
##   (dbl), distance (dbl), hour (dbl), minute (dbl)
```

We would like to collect some information about arrival delays of United Airlines (UA) flights. Do the following: pick all UA departures with non-zero arrival delay, calculate the mean arrival delay for each of the corresponding flight numbers, and find the five flights with the largest mean delay.

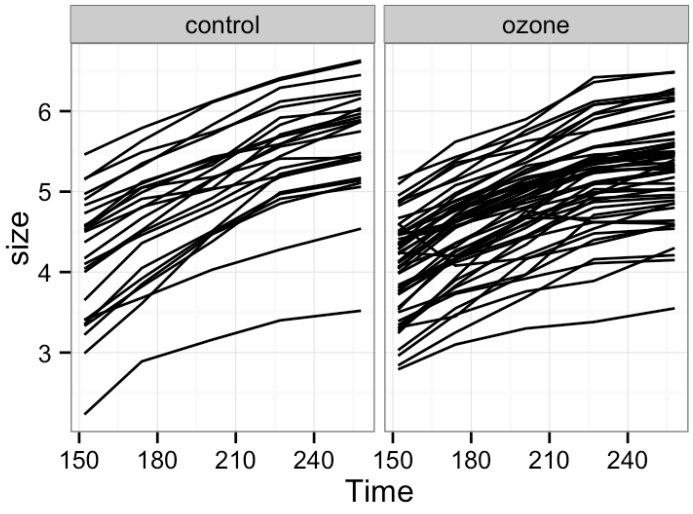
```
flights %>% filter(carrier == "UA" & arr_delay != 0) %>% group_by(flight) %>% summarize(me
an.flight.delay = mean(arr_delay)) %>% arrange(desc(mean.flight.delay)) %>% slice(1:5)
```

```
## Source: local data frame [5 x 2]
##
##   flight mean.flight.delay
## 1   1510             283.0
## 2    125             113.0
## 3    640             111.0
## 4   1084              86.0
## 5    348              85.5
```

Summary of finding goes here.

Question 4: (2 pts) *In an in-class exercise, we made the following plot of the Sitka dataset:*

```
ggplot(Sitka, aes(x=Time, y=size, group=tree)) + geom_line() + facet_wrap(~treat)
```



Now modify the plot so that the line for each tree is colored according to the maximum size of the tree.

```
Sitka %>% group_by(tree) %>% arrange(desc(size)) %>% mutate(max.size = max(size)) -> Sitka
.max
ggplot(Sitka.max, aes(x=Time, y=size, group=tree, color = max.size)) + geom_line() + facet
_wrap(~treat)
```