Homework 4

Manav Mandhani, mm58926

This homework is due on Feb 17, 2015 in class.

Question 1: (2 pts) *For the `msleep` data set from ggplot2, calculate how many distinct orders there are within each "vore". Hint: use the functions `distinct()` and `tally()`.*

```
msleep %>% select(vore) %>% distinct() %>% tally()
```

```
##   n
## 1 5
```

There are 5 distinct orders

Question 2: (4 pts) *Invent two simple data sets that allow you explain the difference between the dplyr functions `left_join()` and `inner_join()`. Explain which features of your data sets affect the behavior of these two functions.*

```
n = c(2, 3, 5)
s = c("aa", "bb", "cc")
b = c(TRUE, FALSE, TRUE)
df1 = data.frame(n, s)

n = c(2, 3, 5, 7, 9)
s1 = c("aa", "bb", "cc","dd","ee")
b1 = c(TRUE, FALSE, TRUE, FALSE, TRUE)
df2 = data.frame(n, s1, b1)

df2 %>% left_join(df1, c("n")) -> left_df
df2 %>% inner_join(df1, c("n")) -> in_df

left_df
```

```
##   n s1    b1    s
## 1 2 aa  TRUE   aa
## 2 3 bb FALSE   bb
## 3 5 cc  TRUE   cc
## 4 7 dd FALSE <NA>
## 5 9 ee  TRUE <NA>
```

```
in_df
```

```
##   n s1    b1  s
## 1 2 aa  TRUE aa
## 2 3 bb FALSE bb
## 3 5 cc  TRUE cc
```

An inner join, as in in_df, looks for a match in n for both data sets.

A left join, as in left_df, takes in all the rows from the left dataset and then adds on all the matches from the right dataset.

Question 3: (2 pts) *The following code downloads a data set containing information about total international air passengers from 1950 to 1961:*

```
air <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/AirPassengers.csv")
air
```

```
##    Year Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1  1950 112 118 132 129 121 135 148 148 136 119 104 118
## 2  1951 115 126 141 135 125 149 170 170 158 133 114 140
## 3  1952 145 150 178 163 172 178 199 199 184 162 146 166
## 4  1953 171 180 193 181 183 218 230 242 209 191 172 194
## 5  1954 196 196 236 235 229 243 264 272 237 211 180 201
## 6  1955 204 188 235 227 234 264 302 293 259 229 203 229
## 7  1956 242 233 267 269 270 315 364 347 312 274 237 278
## 8  1957 284 277 317 313 318 374 413 405 355 306 271 306
## 9  1958 315 301 356 348 355 422 465 467 404 347 305 336
## 10 1959 340 318 362 348 363 435 491 505 404 359 310 337
## 11 1960 360 342 406 396 420 472 548 559 463 407 362 405
## 12 1961 417 391 419 461 472 535 622 606 508 461 390 432
```

Convert this data set into a table with three columns, one for the year, one for the month, and one for the number of passengers.

```
air %>% gather("month", "number_of_passengers", Jan:Dec) %>% head()
```

```
##   Year month number_of_passengers
## 1 1950   Jan                  112
## 2 1951   Jan                  115
## 3 1952   Jan                  145
## 4 1953   Jan                  171
## 5 1954   Jan                  196
## 6 1955   Jan                  204
```

Question 4: (2 pts) *The* `sleep` *dataset contains amount of extra sleep (in hours) for ten students treated with two different drugs each. The drug treatment is indicated in the* `group` *column:*

```
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

Convert this table into a wide table that has three columns, one for student ID, one for extra sleep under treatment 1, and one for extra sleep under treatment 2.

```
sleep %>% spread(group, extra) %>% head()
```

```
##   ID    1    2
## 1  1  0.7  1.9
## 2  2 -1.6  0.8
## 3  3 -0.2  1.1
## 4  4 -1.2  0.1
## 5  5 -0.1 -0.1
## 6  6  3.4  4.4
```