Homework 5

MTH 3270 Data Science

Due Wed., Mar. 11

Read These Chapters of the Book	**Then Do These Exercises**
4 5	Problem 1 (below), Problem 2 (below), Problem 3 (below)*, 4.4*, 4.6*, 4.7* 5.6**, 5.7***

* **Problems 3 (below)**, and **4.4**, **4.6**, and **4.7** (from the book) all use the `"nycflights13"` package, but in addition to the `flights` data, they also use the `planes` and `weather` data sets. Type `?planes` and `?weather` for more info.

** For **Problem 5.6**, you can create the data frame `ds1` using:

```
ds1 <- data.frame(id = rep(1:3, times = 2),
                  group = rep(c("T", "C"), each = 3),
                  vals = c(4, 6, 8, 5, 6, 10))
```

*** For **Problem 5.7**, you can create the data frame using:

```
my.data <- data.frame(grp = rep(c("A", "B"), each = 2),
                      sex = rep(c("F", "M"), times = 2),
                      meanL = c(0.22, 0.47, 0.33, 0.55),
                      sdL = c(0.11, 0.33, 0.11, 0.31),
                      meanR = c(0.34, 0.57, 0.40, 0.65),
                      sdR = c(0.09, 0.33, 0.07, 0.27))
```

1 Consider the following data on houses for sale (from **pg 121** of our textbook ***Modern Data Science with R***):

```
myURL <- "http://tiny.cc/dcf/houses-for-sale.csv"
Houses <- read.csv(myURL)
```

We'll use a *subset* of the variables, namely `fuel`, `heat`, `sewer`, and `construction`:

```
Houses_small <- select(Houses, fuel, heat, sewer, construction)
```

To *recode* `fuel` as `"gas"`, `"electric"`, etc., `sewer` as `"none"`, `"private"`, etc., and so on, we first create a *codebook* data frame that can be used to translate the **integers** to `"character"`:

```
Translations <- read.csv("http://tiny.cc/dcf/house_codes.csv",
                         stringsAsFactors = FALSE)
```

The same information can also be presented in a wide format:

```
CodeVals <- Translations %>% spread(key = system_type,
                                    value = meaning,
                                    fill = "invalid")
```

As an example, below we use `left_join()` to merge `Houses_small` with `CodeVals`, matching rows in `CodeVals` by `code` to rows in `Houses_small` by `fuel`:

```
Houses_small <- left_join(x = Houses_small,
                          y = select(CodeVals, code, fuel_type),
                          by = c(fuel = "code"))
```

Here's the resulting data set, with the *recoded* `fuel` variable:

```
head(Houses_small)
```

a) Report R commands that *recode* the remaining variables in `Houses_small` (`heat`, `sewer`, `construction`), then *remove* the original (**integer**-valued) variables. You should end up with this:

```
head(Houses_small)

##   fuel_type heat_type sewer_type new_const
## 1  electric  electric    private        no
## 2       gas hot water    private        no
## 3       gas hot water     public        no
## 4       gas   hot air    private        no
## 5       gas   hot air     public       yes
## 6       gas   hot air    private        no
```

b) Now (using `Houses_small` obtained in Part *b*), describe in words what the following commands do. Then rewrite them into a more readable version using the **pipe operator %>%**.

```
arrange(summarize(group_by(select(filter(Houses_small, new_const == "no"),
  fuel_type, heat_type), fuel_type), count = n()), desc(count))
```

Hint: Recall that when two function calls are nested, R evaluates the inner one first.

2 Using the `flights` data set (from the `"nycflights13"` package), for each destination (`dest`), determine the *total* minutes of delay and the *average* minutes of delay. Report your R command(s).

3 The `flights` data set contains information about each *flight* in 2013. The `planes` data set contains information about each *airplane*.

a) Which variable would be the ***key*** for combining the two data frames using one of the `*_join()` functions?

b) Combine the `flights` and `planes` data sets using an appropriate `*_join()` function. Which `manufacturer` made the most flights in 2013? How many flights did it make?