Homework #2

Answer the following questions in a single R script called `hw02.R`. Answers must be given by R commands. You cannot simply look at the data set and answer the question via direct inspection. Use comments (#) to indicate which portion of your code answers which question. Be sure that you obtain the correct solutions to each question when you execute your script one line at a time from top to bottom.

Each question will be graded out of 4 points according to the following criteria:

- 0 points: No attempts is made to answer the question.
- 1 point: An attempt is made that, although unsuccessful, revealed some understanding of what the question was asking.
- 2 points: Solution is incorrect, but with some modifications, could be corrected.
- 3 points: Solution is incorrect, but easily resolved with minor modifications **OR** solution is correct, but obtained via convoluted reasoning or by avoiding standard approaches.
- 4 points: Solution is correct and uses standard approaches.

For the following problems, you will use the data contained in `hw02_batting.csv`. This data set contains the batting records for various players in major league baseball.

#1) Give a scatter plot where each point represents a player that played in the 2000 season, with x-coordinate corresponding to the number of "at-bats" (abbreviated `AB`) and the y-coordinate corresponding to the number of "home runs" (abbreviated `HR`).

#2) Major League Baseball actually consists of two separate leagues: the American League (`lgID=AL`) and the National League (`lgID=NL`). Give the same scatter plot as in #1, but now color the points according to which league the player played in.

#3) Repeat the plot in #2, but give a facet plot by year from 2000 to 2009.

#4) Give box plots for each league that represents the number of hits (abbreviated `H`) by players that had over 100 hits.

For the following problems, you will use the data contained in `hw02_exports.csv`. This data set contains a detailed list of U.S. exports in 2006 for various commodities.

#5) Give a bar chart for the value of wheat exported to each continent.

#6) The `end_use_code` column in this data set is used to identify commodities. An `end_use_code` of the form `11***` gives energy sector commodities. We restrict our attention to these commodities. Give a stacked bar chart where each bar corresponds to a continent, the height of the bar corresponds to the total value of these commodities, and each sub-bar is colored according to the value of `commodity` (i.e. the bar is broken up into sub-bars where each sub-bar corresponds to the value of commodities such as "fuel oil" and "natural gas").

#7) We restrict our attention to commodities with end use codes of the form `41***`. Give a fill bar chart where each bar corresponds to such a commodity with the bar colored by `continent`.

#8) Included in `tidyverse` is a function called `map_data`. The following command stores a tibble that contains the latitude and longitude of various countries:

```
1 world <- as_tibble(map_data("world"))
```

With this data, we can then use `geom_map` to create a world map with the color of each country determined by its latitude.

```
ggplot(data=world) +
  geom_map(mapping=aes(map_id=region,fill=lat),map=world) +
  expand_limits(x=world$long,y=world$lat)
```

In the next example, `my_data` contains each region in `world` and assigns a random number between 0 and 1 to it in the column `random`. The map is then colored by region according to the number assigned in `random`.

```
uniq_regions <- unique(world$region)
my_data <- tibble(region=uniq_regions,random=runif(length(uniq_regions)))

ggplot(data=my_data) +
  geom_map(mapping=aes(map_id=region,fill=random),map=world) +
  expand_limits(x=world$long,y=world$lat)
```

Create a map plot that shows the total value of exports sent to each country. You will need to use the `summarize` function to get this correct.