EDA Activity

CMSC320

October 3, 2016

Let's practice some EDA work. We're using the `Wage` dataset provided by the `ISLR` package.

```r
library(tibble)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

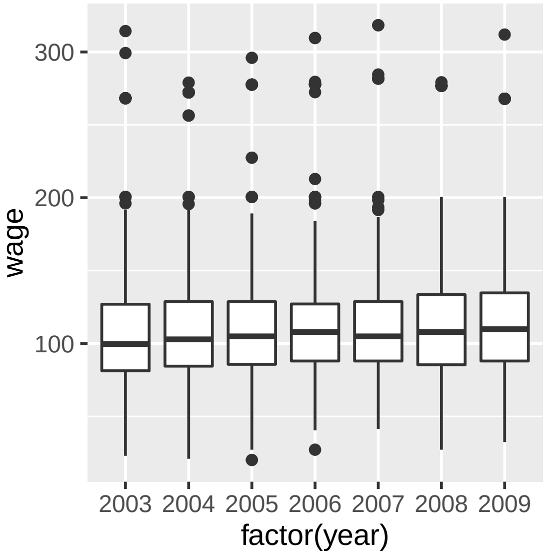
```r
library(ggplot2)

library(ISLR)
data(Wage)

wage <- as_tibble(Wage)
wage
```

```
## # A tibble: 3,000 × 12
##     year   age     sex           maritl     race       education
## *  <int> <int>  <fctr>           <fctr>   <fctr>          <fctr>
## 1   2006    18 1. Male 1. Never Married 1. White    1. < HS Grad
## 2   2004    24 1. Male 1. Never Married 1. White 4. College Grad
## 3   2003    45 1. Male       2. Married 1. White 3. Some College
## 4   2003    43 1. Male       2. Married 3. Asian 4. College Grad
## 5   2005    50 1. Male      4. Divorced 1. White      2. HS Grad
## 6   2008    54 1. Male       2. Married 1. White 4. College Grad
## 7   2009    44 1. Male       2. Married 4. Other 3. Some College
## 8   2008    30 1. Male 1. Never Married 3. Asian 3. Some College
## 9   2006    41 1. Male 1. Never Married 2. Black 3. Some College
## 10  2004    52 1. Male       2. Married 1. White      2. HS Grad
## # ... with 2,990 more rows, and 6 more variables: region <fctr>,
## #   jobclass <fctr>, health <fctr>, health_ins <fctr>, logwage <dbl>,
## #   wage <dbl>
```

Let's warmup with one question:

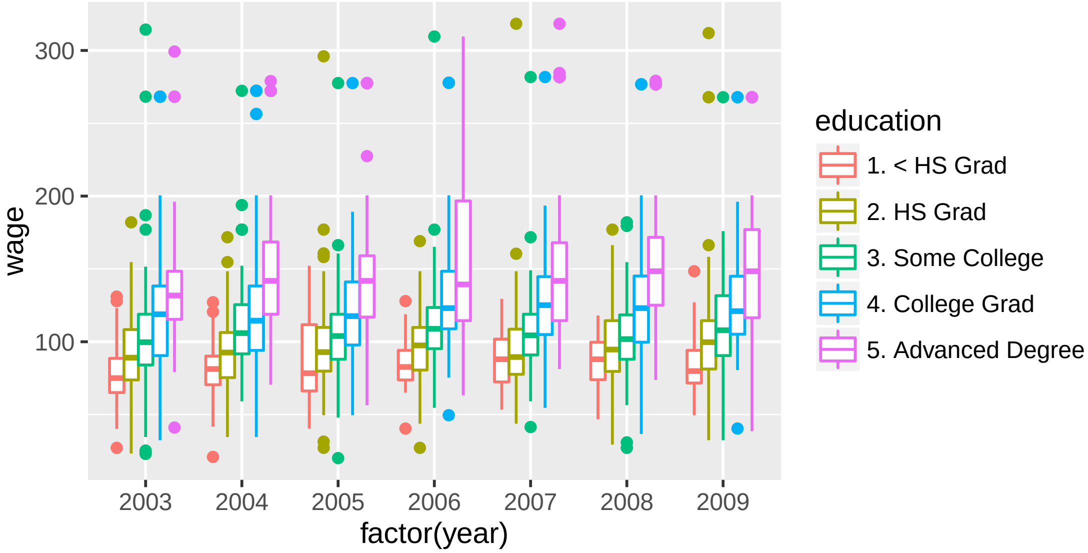
Q0: How are wages distributed overall across years?

```
wage %>%
  ggplot(aes(x=factor(year), y=wage)) +
    geom_boxplot()
```



Now, on your own:

Q1: How are wages distributed across years as a function of education? (Write the code to make this plot)



Q2: How is the central tendency (e.g., median) of wage changing across years?

Q3: How is median wage changing across years as a function of education?

Q4: Is the wage gap between those with advanced degrees and those with less than a HS education changing over time?

Part 1: How are you going to define the wage gap?

Part 2: Make a data frame with columns `year` and `wage_gap`.

Part 3: Plot wage gap as a function of year.