EDA Exercise 5

Aaron Niskin

September 8, 2016

A data set comprised of salaries for professors at an unnamed college can be found at

https://vincentarelbundock.github.io/Rdatasets/csv/car/Salaries.csv

A codebook for the data can be found at

https://vincentarelbundock.github.io/Rdatasets/doc/car/Salaries.html

Download the data and aggregate to address the questions

```
data_sex <- read.csv(data_url)
```

1. How does salary depend on rank and sex?

First we should probably go through and check to see the completeness of our data:

```
str(lapply(data_sex, unique))
```

```
## List of 7
##  $ X            : int [1:397] 1 2 3 4 5 6 7 8 9 10 ...
##  $ rank         : Factor w/ 3 levels "AssocProf","AsstProf",..: 3 2 1
##  $ discipline   : Factor w/ 2 levels "A","B": 2 1
##  $ yrs.since.phd: int [1:53] 19 20 4 45 40 6 30 21 18 12 ...
##  $ yrs.service  : int [1:52] 18 16 3 39 41 6 23 45 20 8 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 1
##  $ salary       : int [1:371] 139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ..
```

So, it seems as though the only non-integer columns are the rank, discipline and sex factor columns. None of those three seem to have any ambiguous or obviously null-signifying values, so we can assume that `complete.cases` will give us an accurate level of correctness:

```
sum(complete.cases(data_sex))
```

```
## [1] 397
```

So, we can conclude that our data is complete. Now to answer the question:

```
aggregate(salary ~ rank + sex, data_sex, mean)
```

```
##        rank    sex    salary
## 1 AssocProf Female  88512.80
## 2  AsstProf Female  78049.91
## 3      Prof Female 121967.61
## 4 AssocProf   Male  94869.70
## 5  AsstProf   Male  81311.46
## 6      Prof   Male 127120.82
```

2. How does salary depend on age and rank?

```
quantile(data_sex$yrs.since.phd)
```

```
##   0%  25%  50%  75% 100%
##    1   12   21   32   56
```

```
quantile(data_sex$yrs.since.phd, seq(0, 10)*0.1)
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  1.0  5.0 10.0 13.0 17.4 21.0 25.0 30.0 35.0 40.0 56.0
```

```
yrsSincePHD<- cut(data_sex$yrs.since.phd,
                  breaks=quantile(data_sex$yrs.since.phd, seq(0, 1, 0.25)),
                  include.lowest = TRUE)
aggregate(salary ~ rank + yrsSincePHD, data_sex, mean)
```

```
##        rank yrsSincePHD    salary
## 1 AssocProf      [1,12]  96475.80
## 2  AsstProf      [1,12]  80775.99
## 3      Prof      [1,12] 113488.86
## 4 AssocProf     (12,21]  93698.72
## 5      Prof     (12,21] 122921.64
## 6 AssocProf     (21,32]  87680.29
## 7      Prof     (21,32] 131914.32
## 8 AssocProf     (32,56]  82775.00
## 9      Prof     (32,56] 125909.78
```

Interestingly enough, it seems as though associate professors tend to get paid more the closer they are to graduation. You'll note that I chose to use `yrs.since.phd` as a signifier for age. This is just working with the data at hand. Of the two age-esque categories, this was more representative of the person's age than the other (which represented the years in service at that school).

3. How does salary depend on discipline and sex?

```
yrsSincePHD<- cut(data_sex$yrs.since.phd,
                  breaks=quantile(data_sex$yrs.since.phd, seq(0, 1, 0.25)),
                  include.lowest = TRUE)
aggregate(salary ~ rank + discipline, data_sex, mean)
```

```
##        rank discipline    salary
## 1 AssocProf          A  83061.12
## 2  AsstProf          A  73935.54
## 3      Prof          A 119948.27
## 4 AssocProf          B 101276.39
## 5  AsstProf          B  84593.91
## 6      Prof          B 133393.76
```