Homework 2

Elisabeth R Silver

2/4/2021

```
library(pacman)
p_load(ggplot2, stats, psych, dplyr, jtools, scales,ggExtra,Hmisc)
#aplpack has some conflicts with Big Sur it seems
#but this link has code to use ggplot instead: https://gist.github.com/benmarwick/00772ccea2dd0b0f1745
# load the functions from this Gist:
devtools::source_gist("00772ccea2dd0b0f1745", filename = "000_geom_bag.r")
```

```
## Sourcing https://gist.githubusercontent.com/benmarwick/00772ccea2dd0b0f1745/raw/4495fff228670fc34afff64709e13ee9b9798780/000_geom_bag.r
```

```
## SHA-1 hash of file is b0623ef97ce124536a31de6acd83924f04c5973a
```

```
devtools::source_gist("00772ccea2dd0b0f1745", filename = "001_bag_functions.r")
```

```
## Sourcing https://gist.githubusercontent.com/benmarwick/00772ccea2dd0b0f1745/raw/4efc4e1da717fe46a23894974de9b3ab4d729d3c/001_bag_functions.r
```

```
## SHA-1 hash of file is a22b94a9fca065e134d9cad573508dabbfb8ee7a
```

Professor salaries

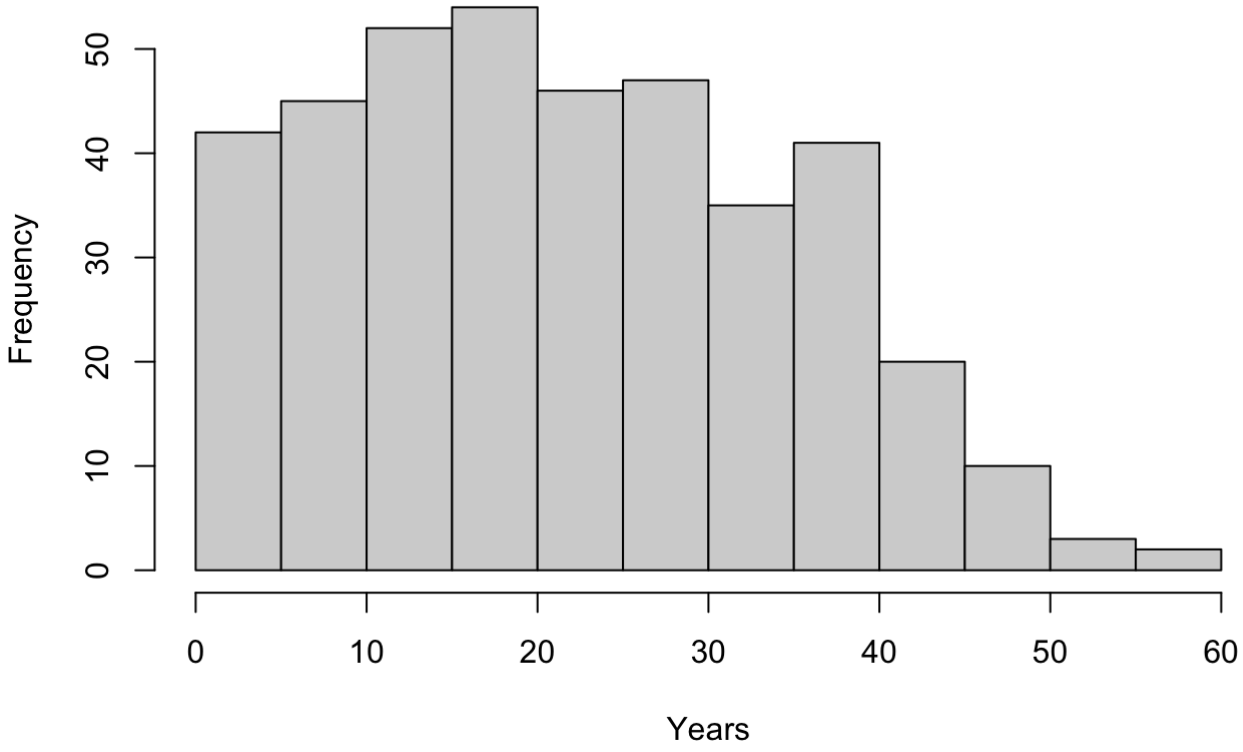
```
df <- carData::Salaries
```

This dataset contains data for professor salaries (*N* = 397) at one instutition for the 2008-2009 school year. I will be focusing on the variables `yrs.since.phd` and `salary`.

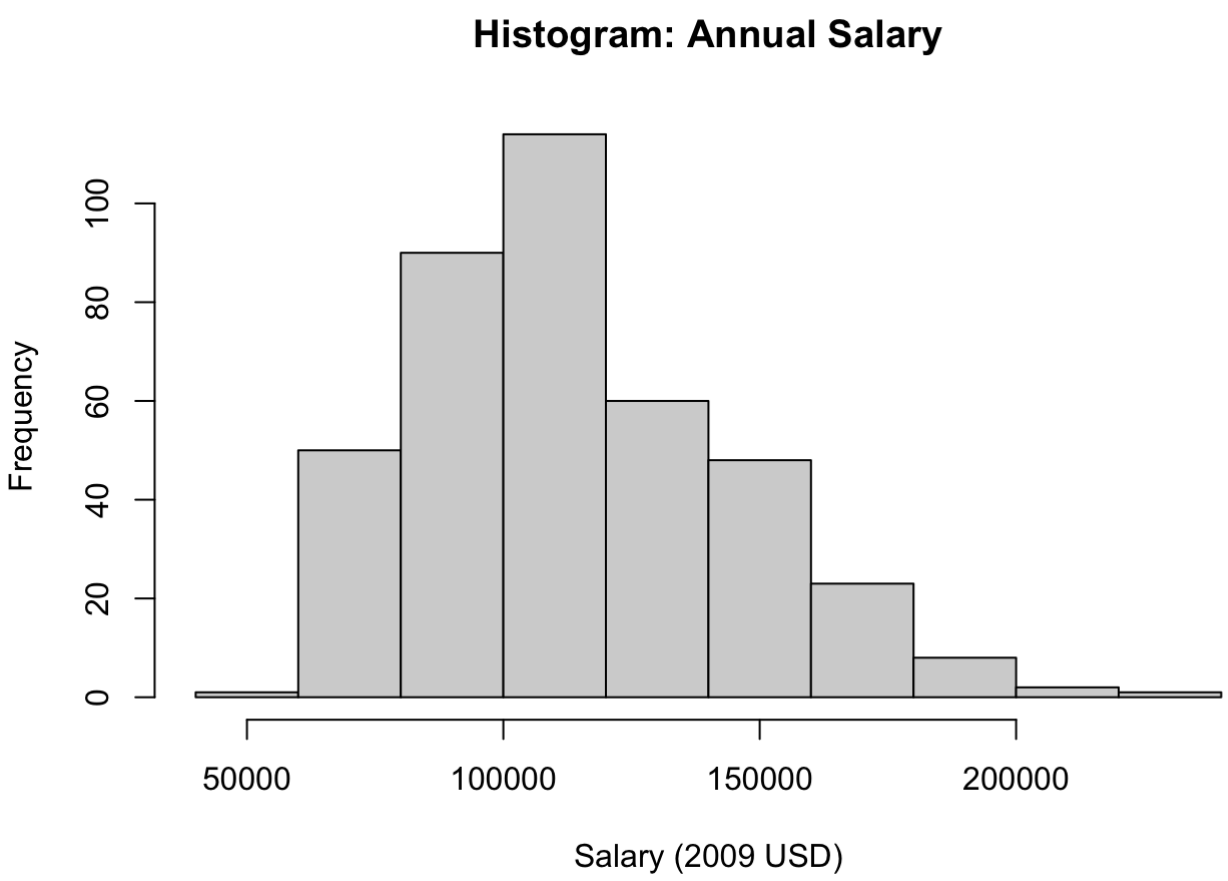
Descriptive statistics

The mean (standard deviation) years since PhD conferral is 22.31 (12.89). The mean (standard deviation) salary is $113,706 ($30,289.04). Both variables are positively skewed.

```
hist(df$yrs.since.phd, main = "Histogram: Years Since PhD Conferral",
     xlab = "Years")
```




```
hist(df$salary, main = 'Histogram: Annual Salary',
     xlab = 'Salary (2009 USD)')
```

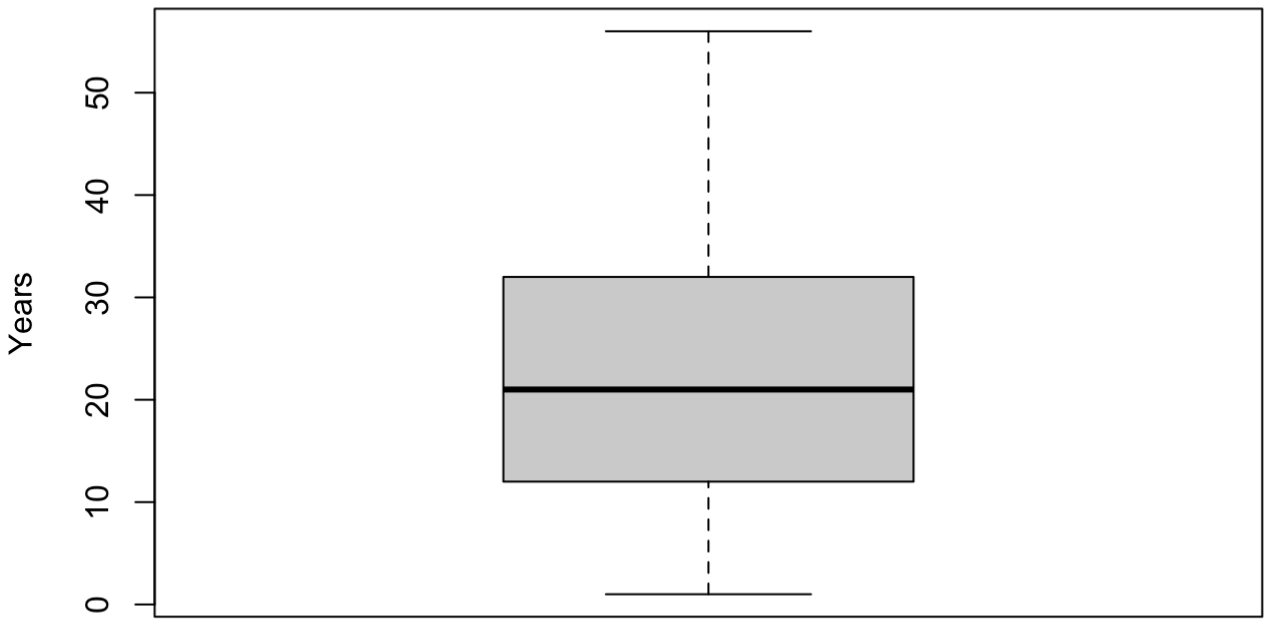


Outliers

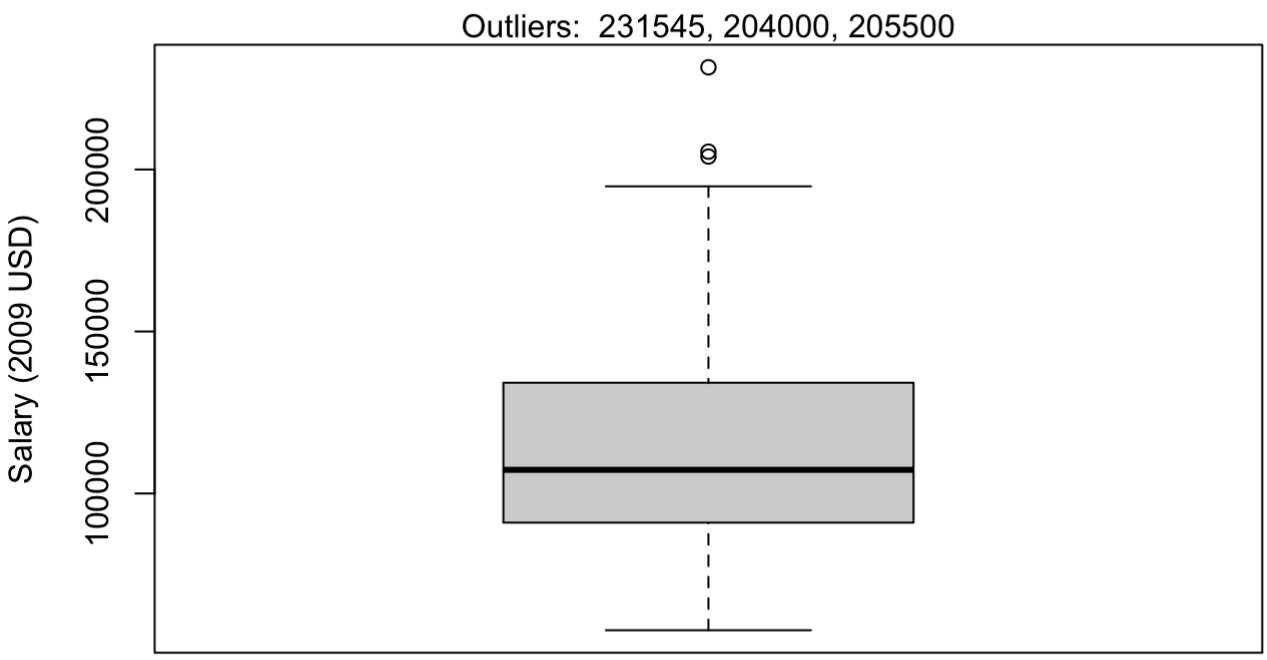
None of the years since PhD conferral are technically outliers, but there seems to be a positive skew.

There do, however, appear to be a few particularly well-paid professors, with three outliers.

```
boxplot(df$yrs.since.phd, main = "Years Since PhD Conferral", ylab = 'Years')
```




```
boxplot(df$salary, main = 'Annual Salary\n', ylab = 'Salary (2009 USD)')
#from https://statsandr.com/blog/outliers-detection-in-r/
out <- boxplot.stats(df$salary)$out
out_ind <- which(df$salary %in% c(out))
mtext(paste("Outliers: ", paste(out, collapse = ", ")))
```



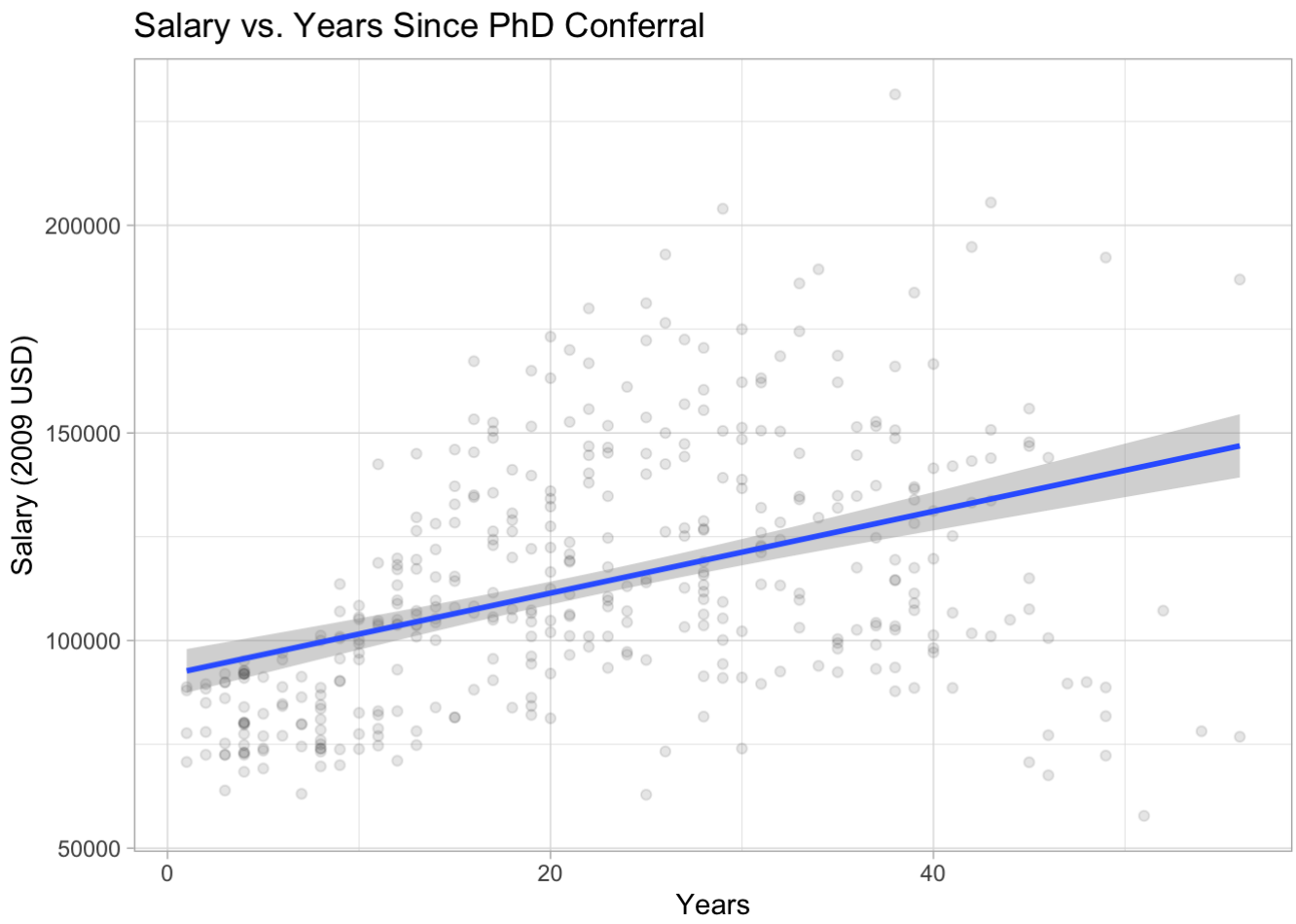
Bivariate exploration

The correlation between salary and years since PhD conferral is 0.42.

It looks like the assumption of linearity is satisfied, but homoscedacity is not.

```
p <- ggplot(df, aes(yrs.since.phd, salary)) +
  geom_point(alpha = 1/10) +
  theme_light() +
  stat_smooth(method=lm)+
  labs(title = "Salary vs. Years Since PhD Conferral", x = "Years", y = "Salary (2009 US
D)")
p
```

```
## `geom_smooth()` using formula 'y ~ x'
```



All of the points are within the fence, so it doesn't look like there are bivariate outliers.

```
p <- ggplot(df, aes(yrs.since.phd, salary, colour=yrs.since.phd)) +
        geom_bag() +
        geom_point(alpha=1/10, show.legend=FALSE) +
        theme_bw()+
        ylab("Salary (2009) USD")+
        xlab("Years since PhD Conferral")
p
```

```
## Warning: In prcomp.default(xydata, na.action = na.omit) :
##  extra argument 'na.action' will be disregarded
```


```
mod <- lm(salary~yrs.since.phd, data = df)
summary(mod)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84171 -19432  -2858  16086 102383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    91718.7     2765.8  33.162   <2e-16 ***
## yrs.since.phd    985.3      107.4   9.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27530 on 395 degrees of freedom
## Multiple R-squared:  0.1758, Adjusted R-squared:  0.1737
## F-statistic: 84.23 on 1 and 395 DF,  p-value: < 2.2e-16
```

The adjusted R^2 suggests that 17.4% of the variance in salary can be accounted for by years since PhD conferral. The estimate for the intercept, 91,718.70, suggests that if the regression line were drawn to cross the y-axis (in other words, someone had 0 years since their PhD was conferred), their salary would be, on average, $91,718.70

(which seems a bit… optimistic). The slope is 985.30, suggesting that for every year following PhD conferral, one's salary increases by $985.30 on average. The p-values are < 0.001, suggesting that the probability that these estimates reflect mere chance is very low.