Lab 5: Wages

JAKE JAMES

February 17, 2021

Introduction

Consider the dataset `Wages1` from the `Ecdat` package.

```
##   exper    sex school     wage
## 1     9 female     13 6.315296
## 2    12 female     12 5.479770
## 3    11 female     11 3.642170
## 4     9 female     14 4.593337
## 5     8 female     14 2.418157
## 6     9 female     14 2.094058
```

This observational dataset records the years experienced, the years schooled, the sex, and the hourly wage for 3,294 workers. *A Guide to Modern Econometrics* by Marno Verbeek utilizes this data in a linear regression context. According to Marno Verbeek, this data is a subsample from the US National Longitudinal Study.

The purpose of this lab is to practice the creative process in exploratory data analysis of asking questions and then investigating those questions using visuals and statistical summaries. It is your job to apply your detective skills to the information hidden in this data. For future use, utilize the modified dataset `wage` according to the R code below:

```
wage=as.tibble(Wages1) %>%
  rename(experience=exper) %>%
  arrange(school)
```

```
## Warning: `as.tibble()` is deprecated as of tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
head(wage)
```

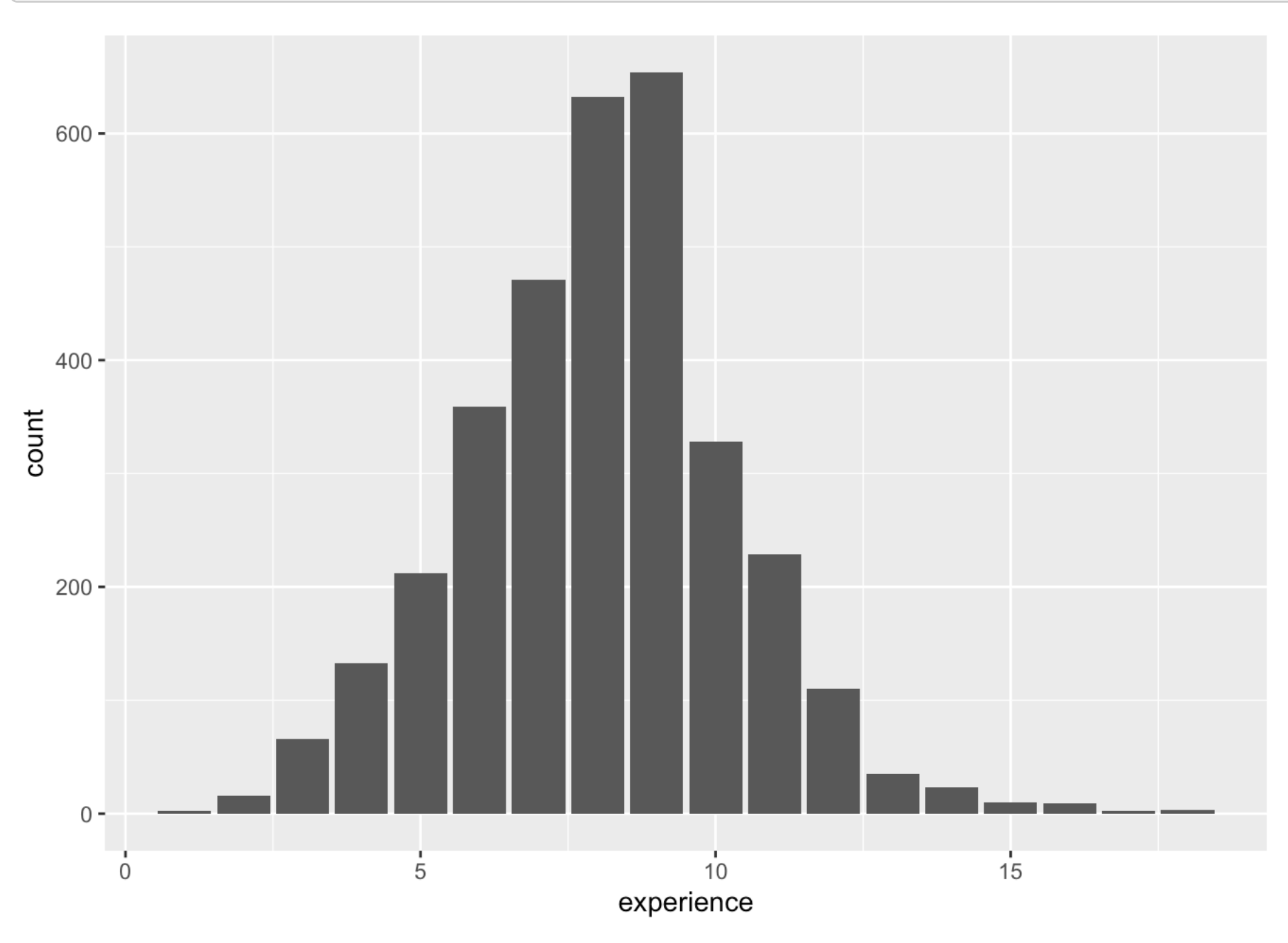
```
## # A tibble: 6 x 4
##   experience sex    school  wage
##        <int> <fct>   <int> <dbl>
## 1         18 male        3 5.52
## 2         15 male        4 3.56
## 3         18 male        4 9.10
## 4         10 female      5 0.603
## 5         11 male        5 3.80
## 6         14 male        5 7.50
```

Part 1: Questions About Variation

Question 1: What is the most common number of years of experience found in the data?

```
## Uses a bar graph to see the distribution of experience in the dataset.

ggplot(wage) +
  geom_bar(aes(x=experience))
```



The most common value for years of experience is 9 and occurs 654 times.

```
wage %>%
  group_by(experience) %>%
  summarize(n=n()) %>%
  arrange(desc(n)) %>%
  summarize(common.exp=first(experience),common.n=first(n))
```

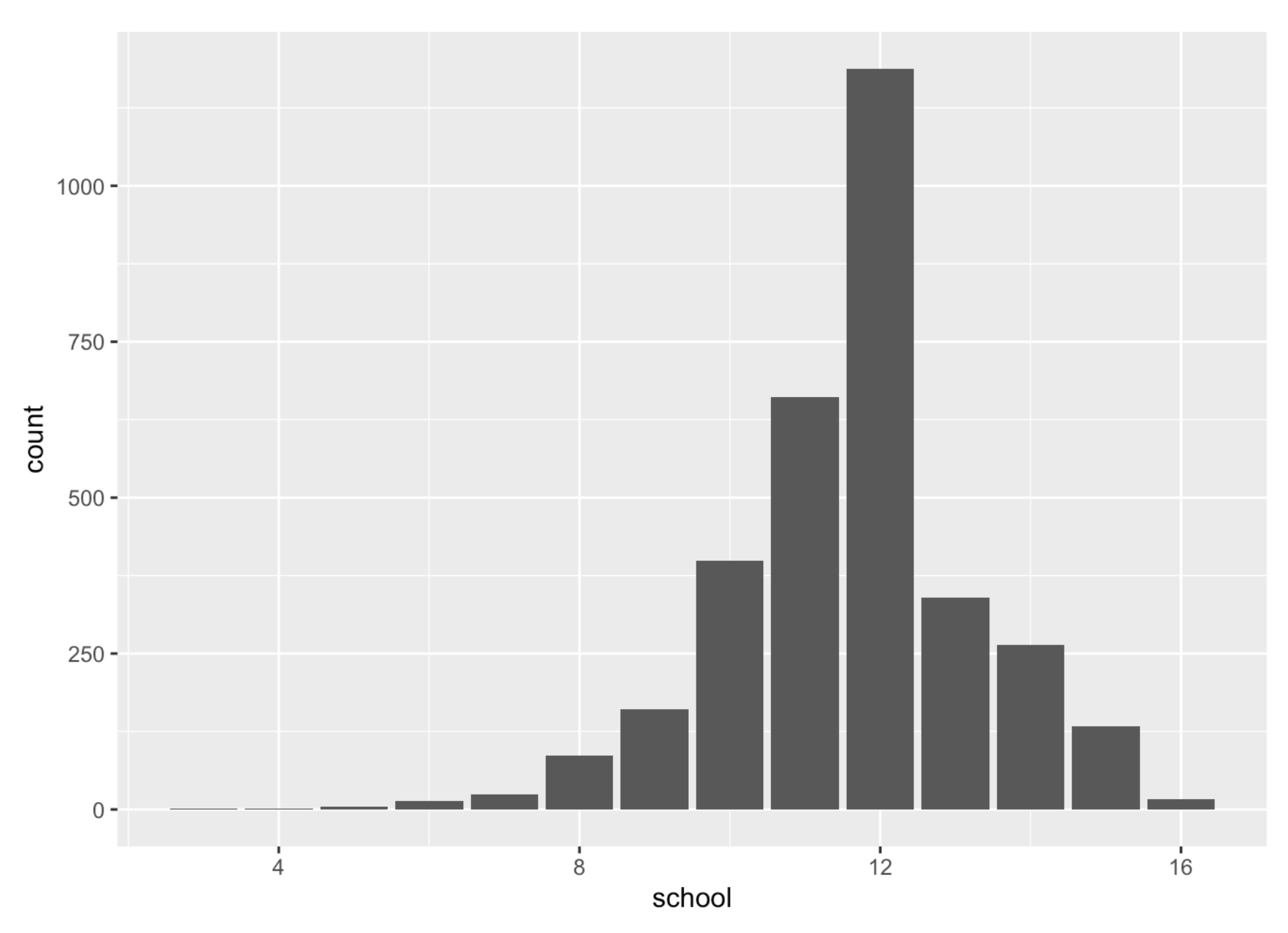
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 2
##   common.exp common.n
##        <int>    <int>
## 1          9      654
```

Question 2: What is the maximum number for years of schooling found in the data?

```
## Uses a bar graph to see the distribution of schooling in the dataset.

ggplot(wage) +
  geom_bar(aes(x=school))
```



The maximum number for years of schooling is 12, and 1,188 employees have this designation.

```
wage %>%
  group_by(school) %>%
  summarize(n=n()) %>%
  arrange(desc(n)) %>%
  summarize(max.school=first(school),
            max.n=first(n))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

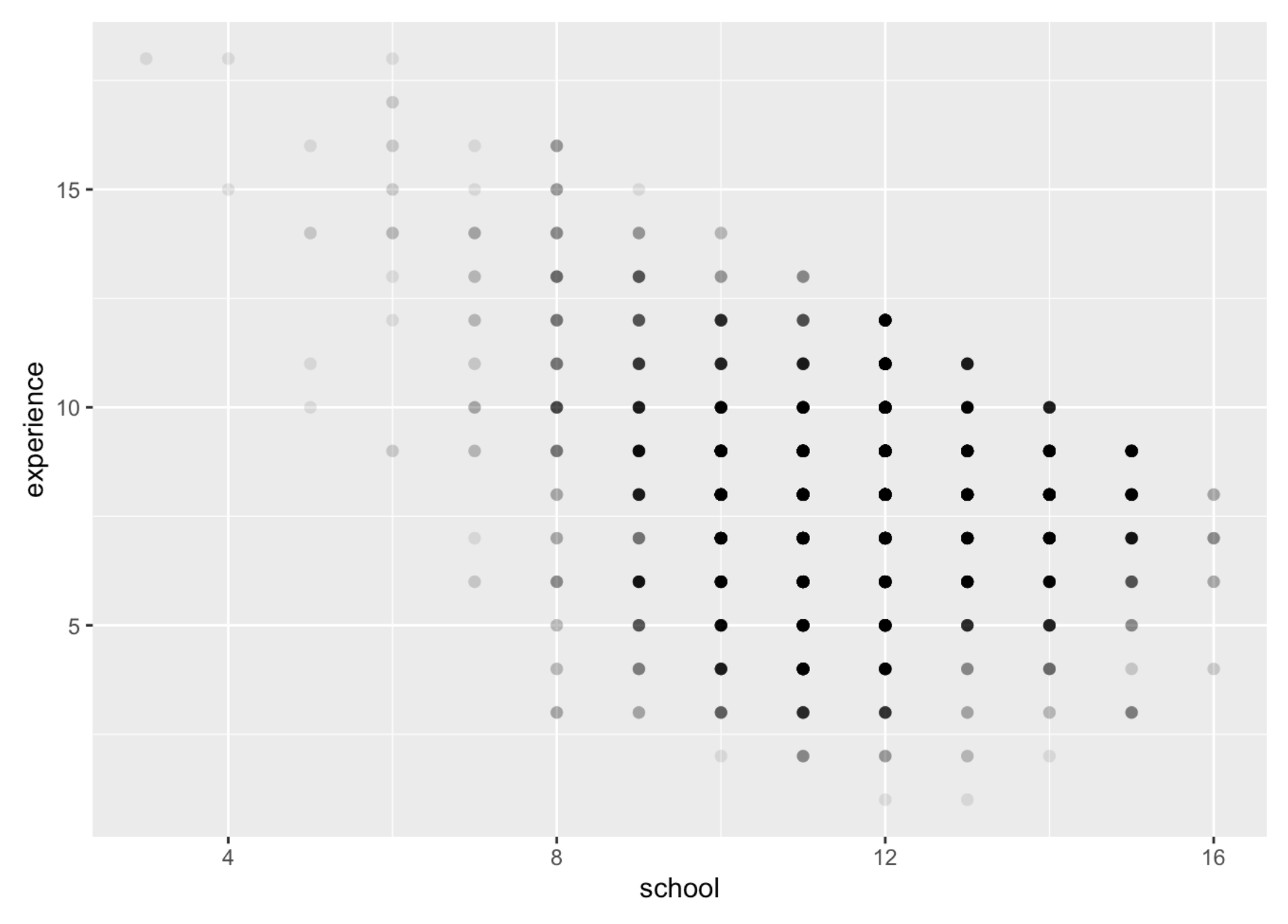
```
## # A tibble: 1 x 2
##   max.school max.n
##        <int> <int>
## 1         12  1188
```

Part 2: Questions about Covariation

Follow-up to Questions 1-2: Is there a relationship between level of schooling and level of experience?

```
## Shows the relationship between schooling and experience.

ggplot(wage) +
  geom_point(aes(x=school,y=experience),
             alpha=0.1,shape=16,size=2)
```



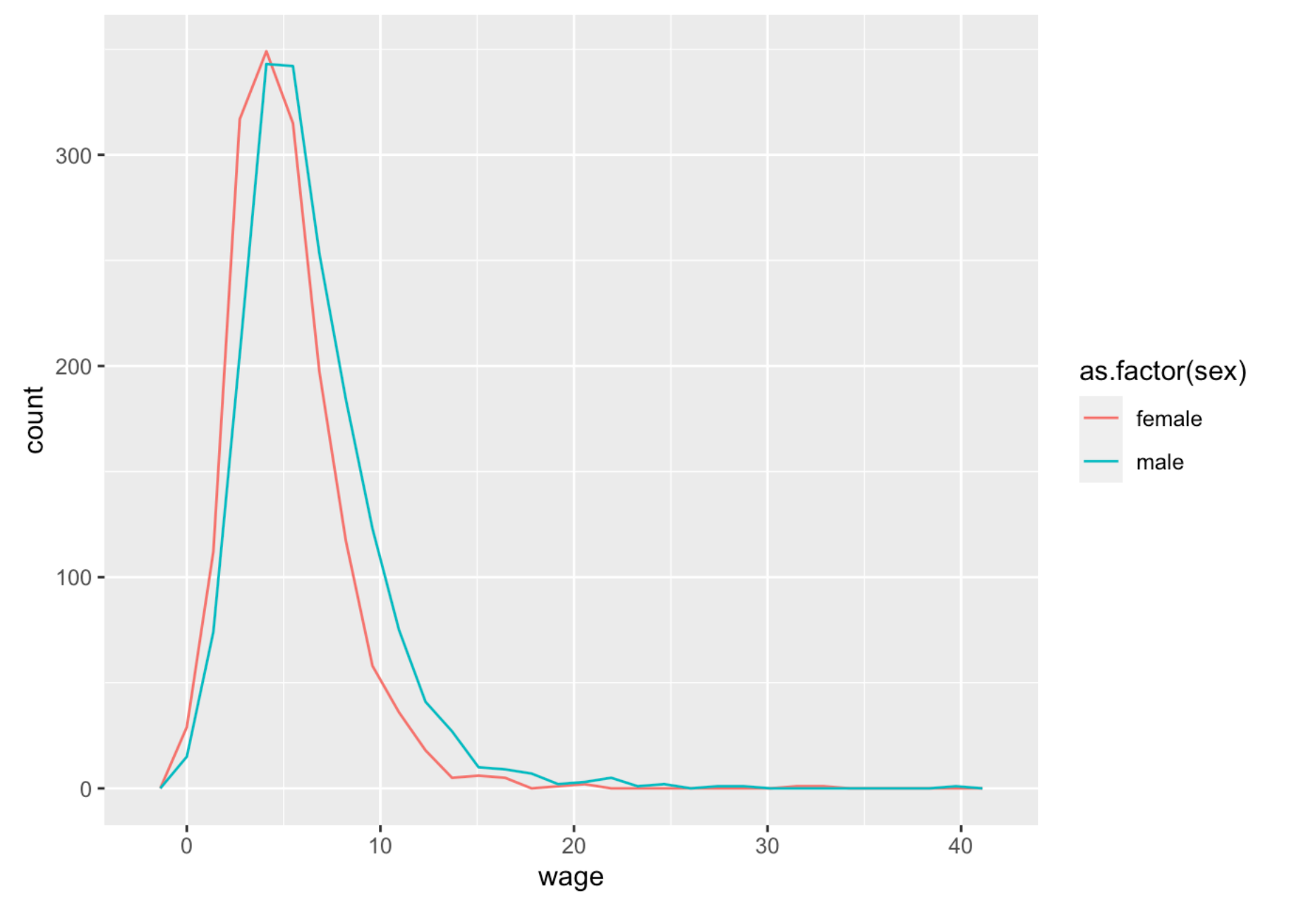
The years of experience seem to decrease as the years of schooling increases.

Possible reason for this: Less schooling = more years to work.

Question 3: How do hourly wages differ between males and females?

```
ggplot(wage)+
  geom_freqpoly(aes(x=wage,color=as.factor(sex)))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



They look similar in wages of 20+ but there are more males in the 5-20 range and more females in the 0-5 range.

```
## Calculates a confidence interval for wages by sex.

wage %>%
  group_by(sex) %>%
  summarize(n=n(),mean=mean(wage),se=sd(wage)/sqrt(n),
            lb=mean-2*se,ub=mean+2*se)
```

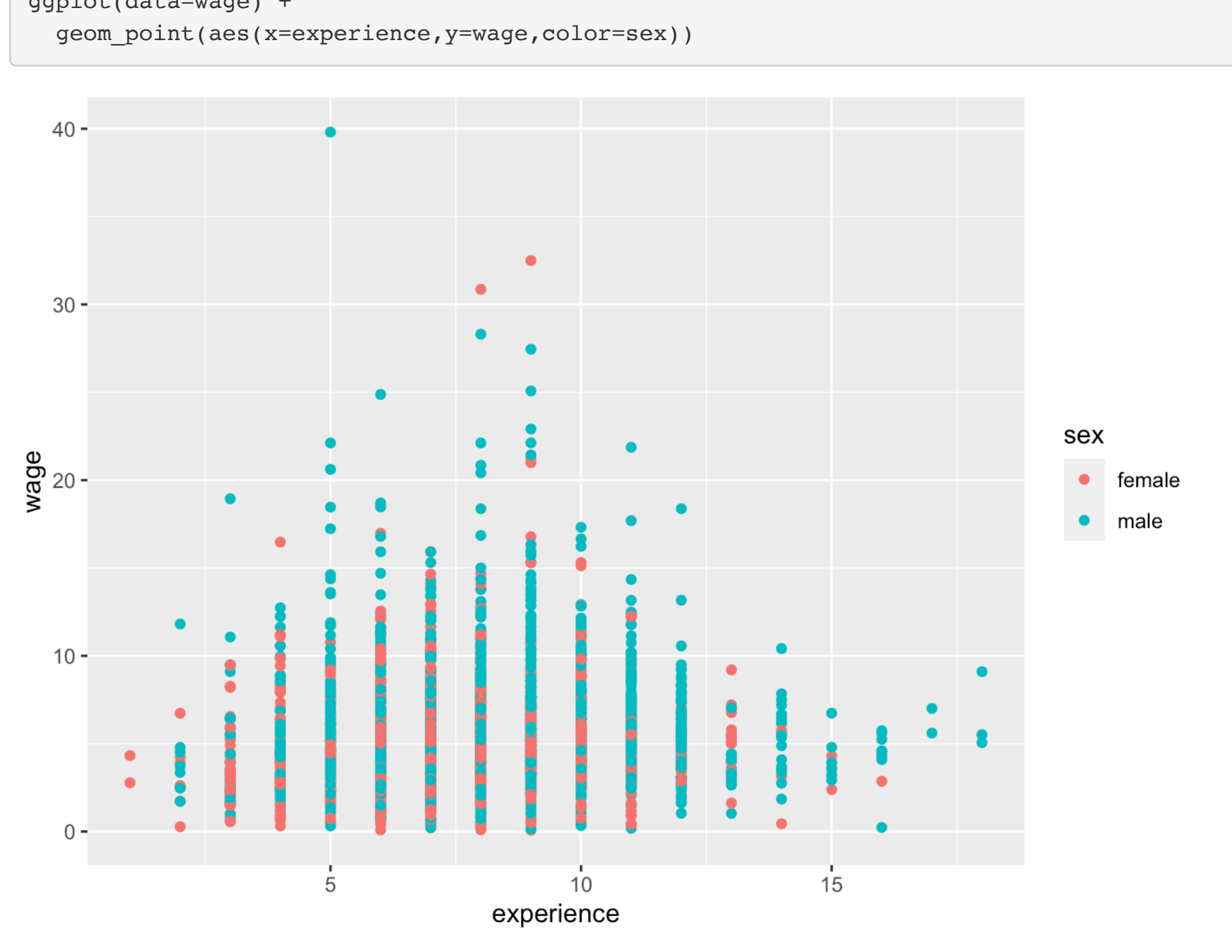
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 6
##   sex        n  mean     se    lb    ub
##   <fct>  <int> <dbl>  <dbl> <dbl> <dbl>
## 1 female  1569  5.15 0.0726  5.00  5.29
## 2 male    1725  6.31 0.0842  6.14  6.48
```

Based on the confidence limits, we have statistical evidence to say that the average hourly wage for men is larger than the average hourly wage for women. We can say with 95% confidence that there is evidence of a significant difference between wages for males and females.

Question 4: Does the relationship between hourly wage and years of experience differ between the sexes?

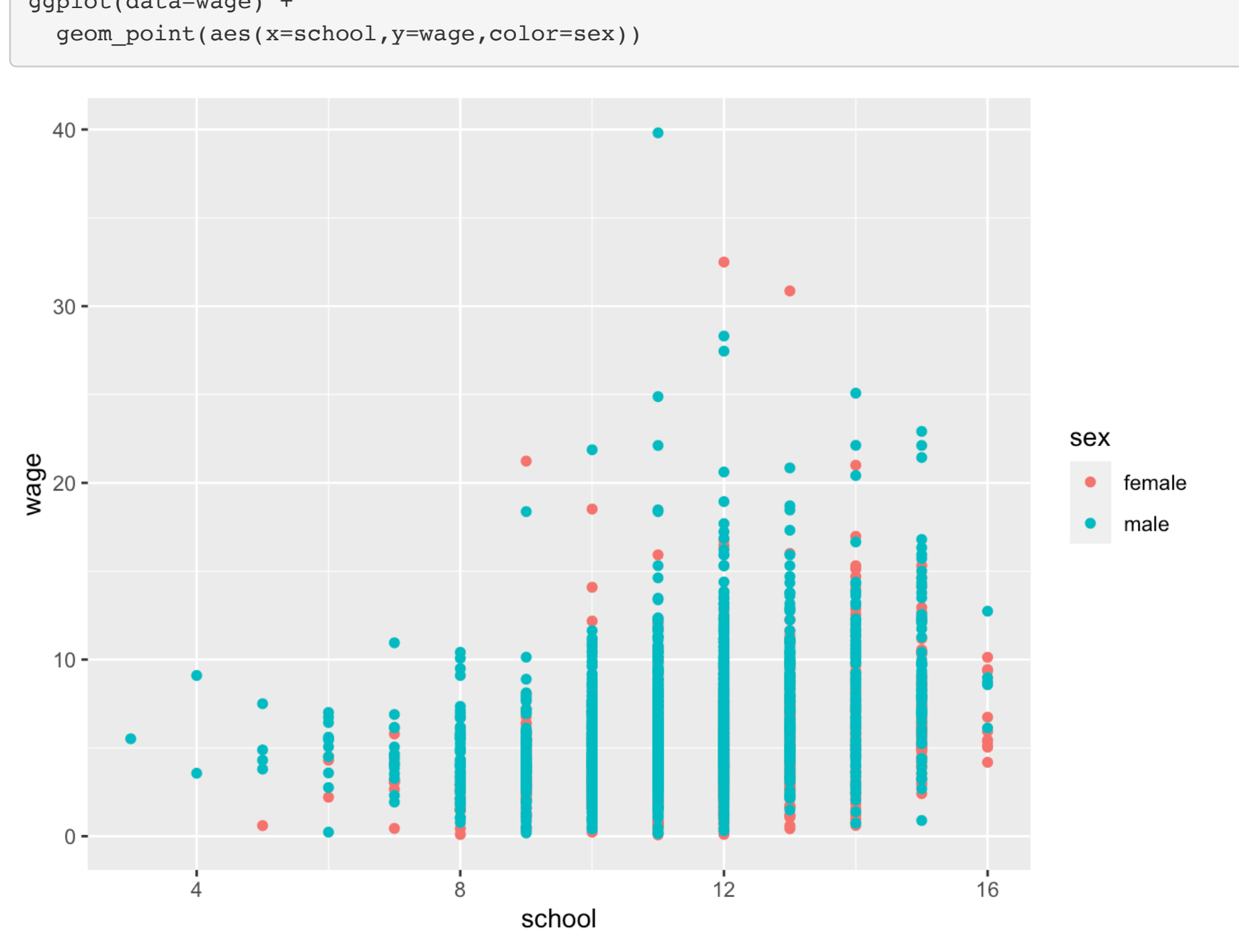
```
ggplot(data=wage) +
  geom_point(aes(x=experience,y=wage,color=sex))
```



There seems to be similar distributions for both sexes, however the males seem to have a higher peak wage for each level of experience.

Question 5: Does the relationship between hourly wage and years of schooling differ between the sexes?

```
ggplot(data=wage) +
  geom_point(aes(x=school,y=wage,color=sex))
```



There doesn't seem to be a clear distinction between female and male regarding the relationship, however it is unclear with the overlay.