# IDS 702: Module 1.1

## Motivating example

Dr. Olanrewaju Michael Akande

# INTRODUCTION

- By now, you should already be familiar with t-tests and simple linear regression (SLR).
- At the very least, you should know the basics.
- Specifically, you should know how to fit a SLR model and assess whether or not the model assumptions are violated.
- We will use those ideas as building blocks for the models we will explore throughout this course.

# MOTIVATING EXAMPLE

- In the 1970's, Harris Trust and Savings Bank was sued for discrimination on the basis of sex.
- As evidence, the defense presented analysis of salaries of employees of one type (skilled, entry level clerical).
- The data is in the file `wagediscrim.txt` on Sakai.
- We are interested in answering the question: **did female employees tend to receive lower base/starting salaries than similarly qualified and experienced male employees?**

Which statistical tests can we use to probe the question above?

# Data

93 employees on data file (61 female, 32 male).

| Variable | Description |
|---|---|
| bsal | Annual salary at time of hire |
| sal77 | Annual salary in 1977. |
| educ | years of education. |
| exper | months previous work prior to hire at bank. |
| fsex | 1 if female, 0 if male |
| senior | months worked at bank since hired |
| age | months |

Since we care about inference on bsal, as our response variable, we will exclude sal77 for all analysis.

Is this reasonable? Why or why not?

# Data

How many rows? How many columns?

```
wages <- read.csv("data/wagediscrim.txt", header= T)
dim(wages)
```

```
## [1] 93  8
```

Take a look at the first few rows of the data.

```
head(wages)
```

```
##   bsal sal77  sex senior age educ exper fsex
## 1 5040 12420 Male     96 329   15  14.0    0
## 2 6300 12060 Male     82 357   15  72.0    0
## 3 6000 15120 Male     67 315   15  35.5    0
## 4 6000 16320 Male     97 354   12  24.0    0
## 5 6000 12300 Male     66 351   12  56.0    0
## 6 6840 10380 Male     92 374   15  41.5    0
```

# Data

Check variable types.

```
wages$sex <- factor(wages$sex,levels=c("Male","Female"))
wages$fsex <- factor(wages$fsex)
str(wages)
```

```
## 'data.frame':    93 obs. of  8 variables:
##  $ bsal  : int  5040 6300 6000 6000 6000 6840 8100 6000 6000 6900 ...
##  $ sal77 : int  12420 12060 15120 16320 12300 10380 13980 10140 12360 10920 ...
##  $ sex   : Factor w/ 2 levels "Male","Female": 1 1 1 1 1 1 1 1 1 1 ...
##  $ senior: int  96 82 67 97 66 92 66 82 88 75 ...
##  $ age   : int  329 357 315 354 351 374 369 363 555 416 ...
##  $ educ  : int  15 15 15 12 12 15 16 12 12 15 ...
##  $ exper : num  14 72 35.5 24 56 41.5 54.5 32 252 132 ...
##  $ fsex  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

# Exploratory data analysis (EDA)

Next, quick summaries of each variable.

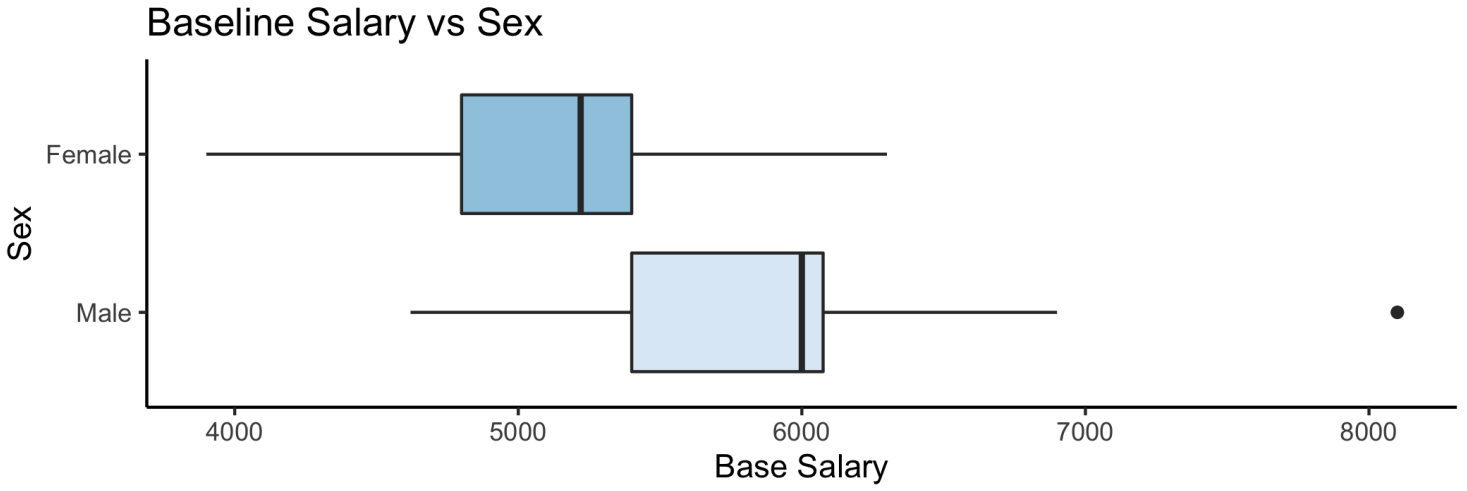
```
summary(wages)
```

```
##       bsal          sal77           sex         senior          age       
##  Min.   :3900   Min.   : 7860   Male  :32   Min.   :65.00   Min.   :280.0  
##  1st Qu.:4980   1st Qu.: 9000   Female:61   1st Qu.:74.00   1st Qu.:349.0  
##  Median :5400   Median :10020               Median :84.00   Median :468.0  
##  Mean   :5420   Mean   :10393               Mean   :82.28   Mean   :474.4  
##  3rd Qu.:6000   3rd Qu.:11220               3rd Qu.:90.00   3rd Qu.:590.0  
##  Max.   :8100   Max.   :16320               Max.   :98.00   Max.   :774.0  
##       educ           exper       fsex   
##  Min.   : 8.00   Min.   :  0.0   0:32  
##  1st Qu.:12.00   1st Qu.: 35.5   1:61  
##  Median :12.00   Median : 70.0         
##  Mean   :12.51   Mean   :100.9         
##  3rd Qu.:15.00   3rd Qu.:144.0         
##  Max.   :16.00   Max.   :381.0
```

# EDA

Since we only care about comparing starting salaries for male and female employees for now, let's look at boxplots of bsal by sex.

```
ggplot(wages,aes(x=sex, y=bsal, fill=sex)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Baseline Salary vs Sex",y="Base Salary",x="Sex") +
   theme_classic() + theme(legend.position="none")
```




What do you think? What can you infer from this plot?

# T-TEST?

We could go further and try a t-test for the hypotheses.

$$H_0 : \mu_{\text{male}} - \mu_{\text{female}} \leq 0 \text{ vs. } H_A : \mu_{\text{male}} - \mu_{\text{female}} > 0$$

```
t.test(bsal~sex,data=wages,alternative="greater")
```

```
##
##     Welch Two Sample t-test
##
## data:  bsal by sex
## t = 5.83, df = 51.329, p-value = 1.855e-07
## alternative hypothesis: true difference in means between group Male and group Female is greater than 0
## 95 percent confidence interval:
##  582.9857      Inf
## sample estimates:
##   mean in group Male mean in group Female
##             5956.875             5138.852
```

Is a t-test sufficient here? Any concerns?

# SLR?

How about fitting a SLR model to the two variables.

$$\text{bsal}_i = \beta_0 + \beta_1 \text{sex}_i + \epsilon_i; \;\; \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

```
model1 <- lm(bsal~sex,data=wages); summary(model1)
```

```
##
## Call:
## lm(formula = bsal ~ sex, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1336.88  -338.85    43.12   261.15  2143.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5956.9      105.3  56.580  < 2e-16
## sexFemale     -818.0      130.0  -6.293 1.08e-08
##
## Residual standard error: 595.6 on 91 degrees of freedom
## Multiple R-squared:  0.3032,    Adjusted R-squared:  0.2955
## F-statistic:  39.6 on 1 and 91 DF,  p-value: 1.076e-08
```

What can we infer from these results?

# EDA

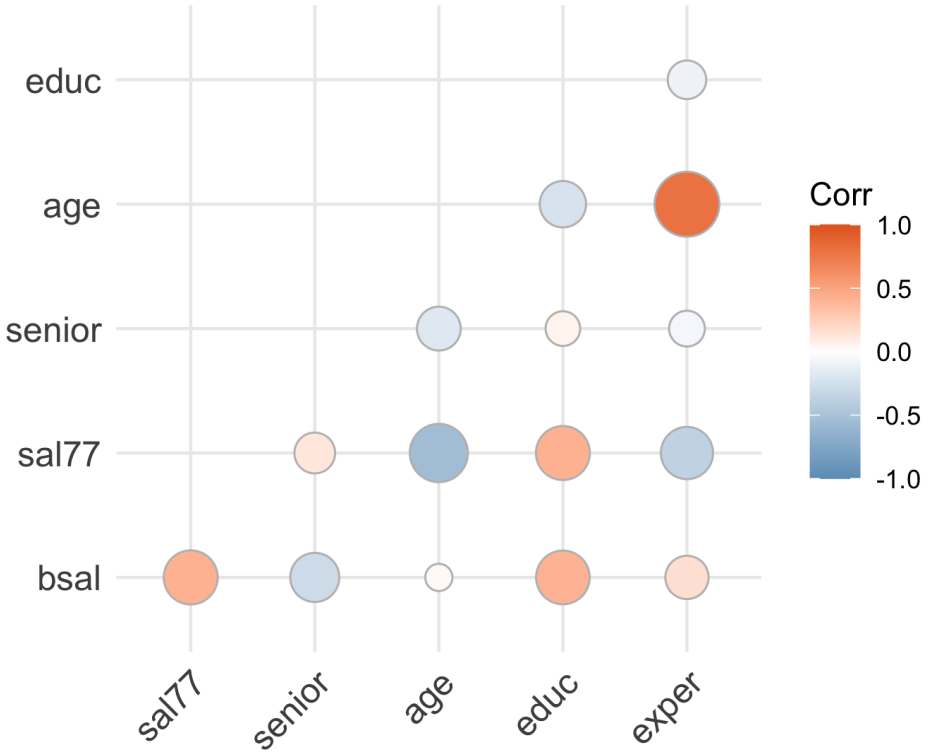
- T-test shows men started at higher salaries than women $(t = 5.83, p < .0001)$; same conclusion from the regression.
- But one could argue this is so because both methods **do not** control for other characteristics. Indeed, we have ignored the other variables.
- There are other variables that are correlated with bsal. Here's the correlation matrix of all numerical variables using the corr function in R.

| | bsal | sal77 | senior | age | educ | exper |
|---|---|---|---|---|---|---|
| bsal | 1.00 | 0.42 | -0.29 | 0.03 | 0.41 | 0.17 |
| sal77 | 0.42 | 1.00 | 0.13 | -0.55 | 0.42 | -0.37 |
| senior | -0.29 | 0.13 | 1.00 | -0.18 | 0.06 | -0.07 |
| age | 0.03 | -0.55 | -0.18 | 1.00 | -0.23 | 0.80 |
| educ | 0.41 | 0.42 | 0.06 | -0.23 | 1.00 | -0.10 |
| exper | 0.17 | -0.37 | -0.07 | 0.80 | -0.10 | 1.00 |

# EDA

Or visually (using the `ggcorrplot` package),

```
wages_corr <- round(cor(wages[,!is.element(colnames(wages),c("sex","fsex"))]),2)
ggcorrplot(wages_corr, method = "circle",type = "lower",
           colors = c("#6D9EC1", "white", "#E46726"))
```



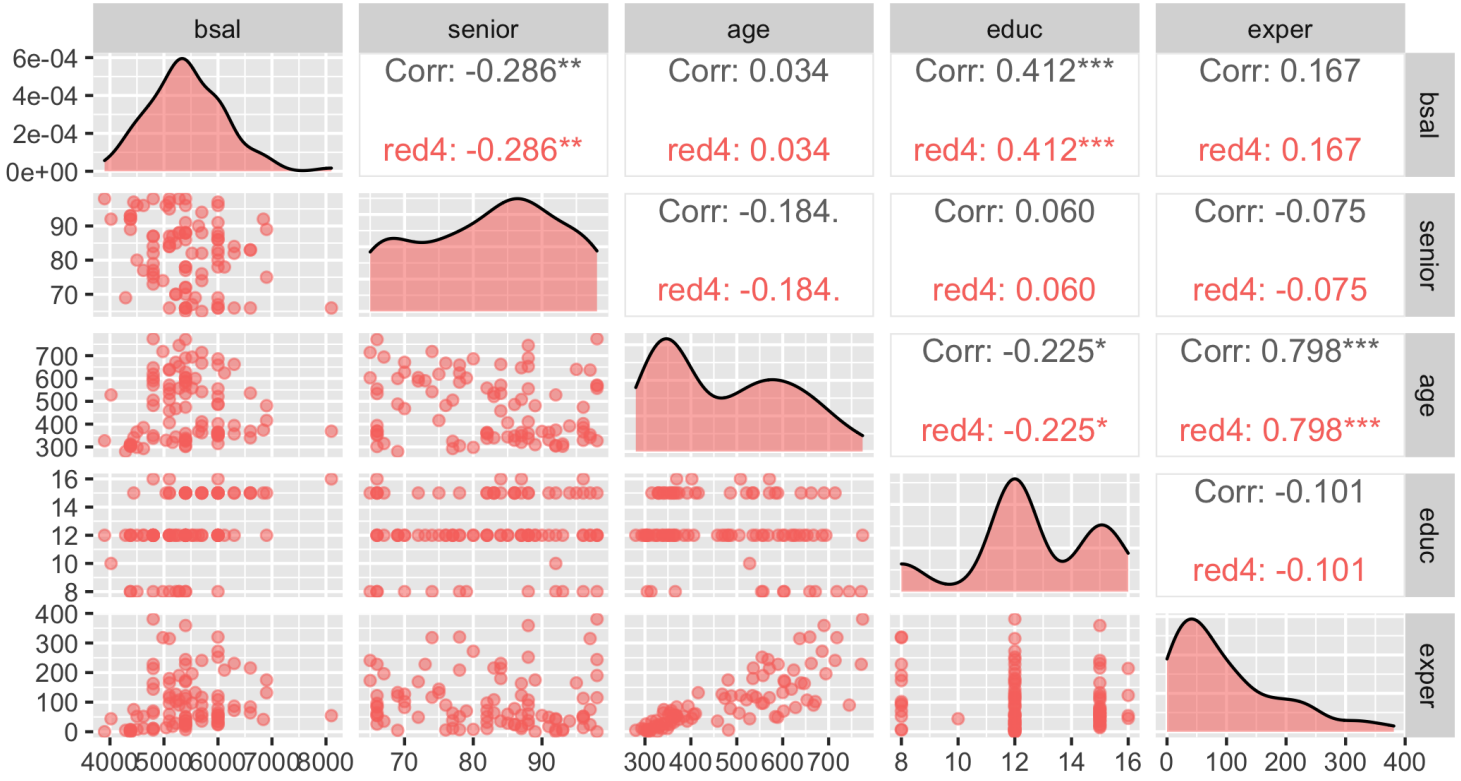
# EDA

- So, let's take a look at scatter plots of all variables
- First, recall the description of all the variables.

| Variable | Description |
|---|---|
| bsal | Annual salary at time of hire |
| sal77 | Annual salary in 1977. |
| educ | years of education. |
| exper | months previous work prior to hire at bank. |
| fsex | 1 if female, 0 if male |
| senior | months worked at bank since hired |
| age | months |

# EDA

```
ggpairs(wages[,!is.element(colnames(wages),c("sal77","sex","fsex"))],
        mapping=ggplot2::aes(colour = "red4",alpha=0.6)) #GGally package
```

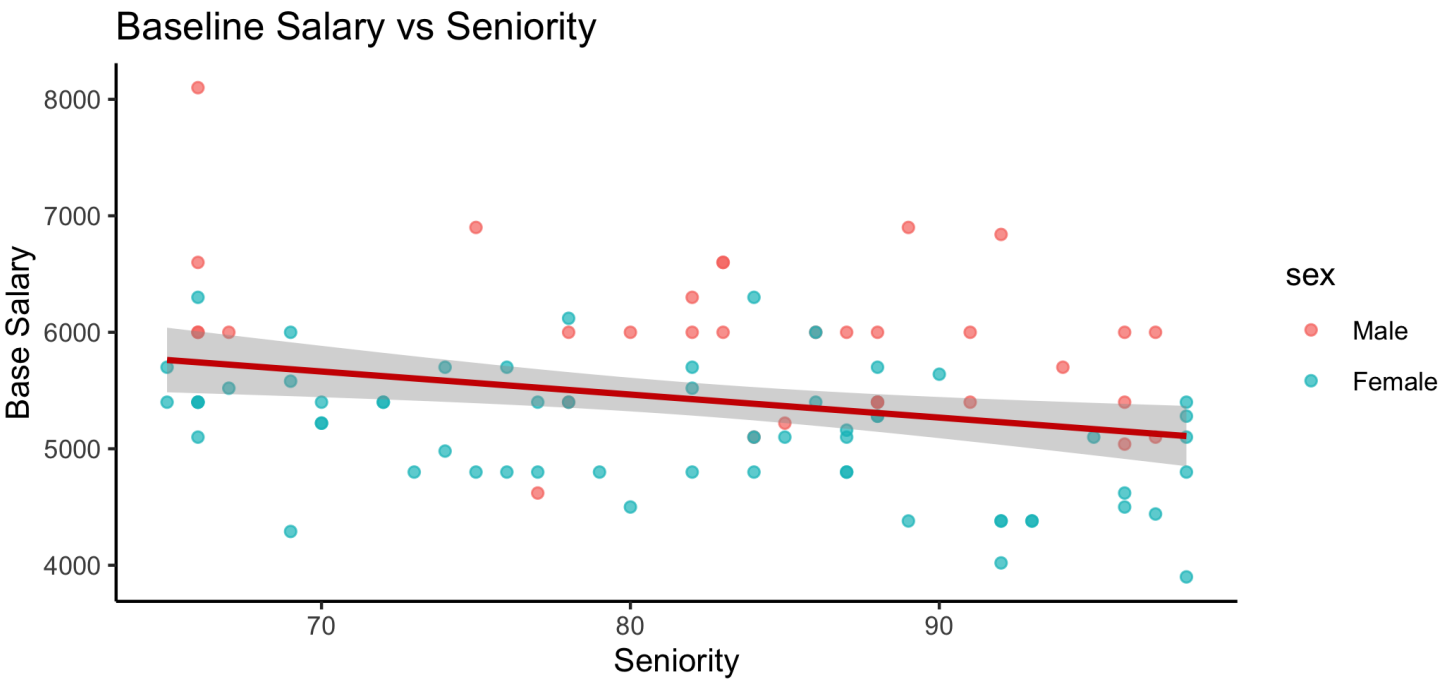



This plot looks very busy!

# EDA

Let's take a closer look one variable at a time. First, bsal vs. senior.

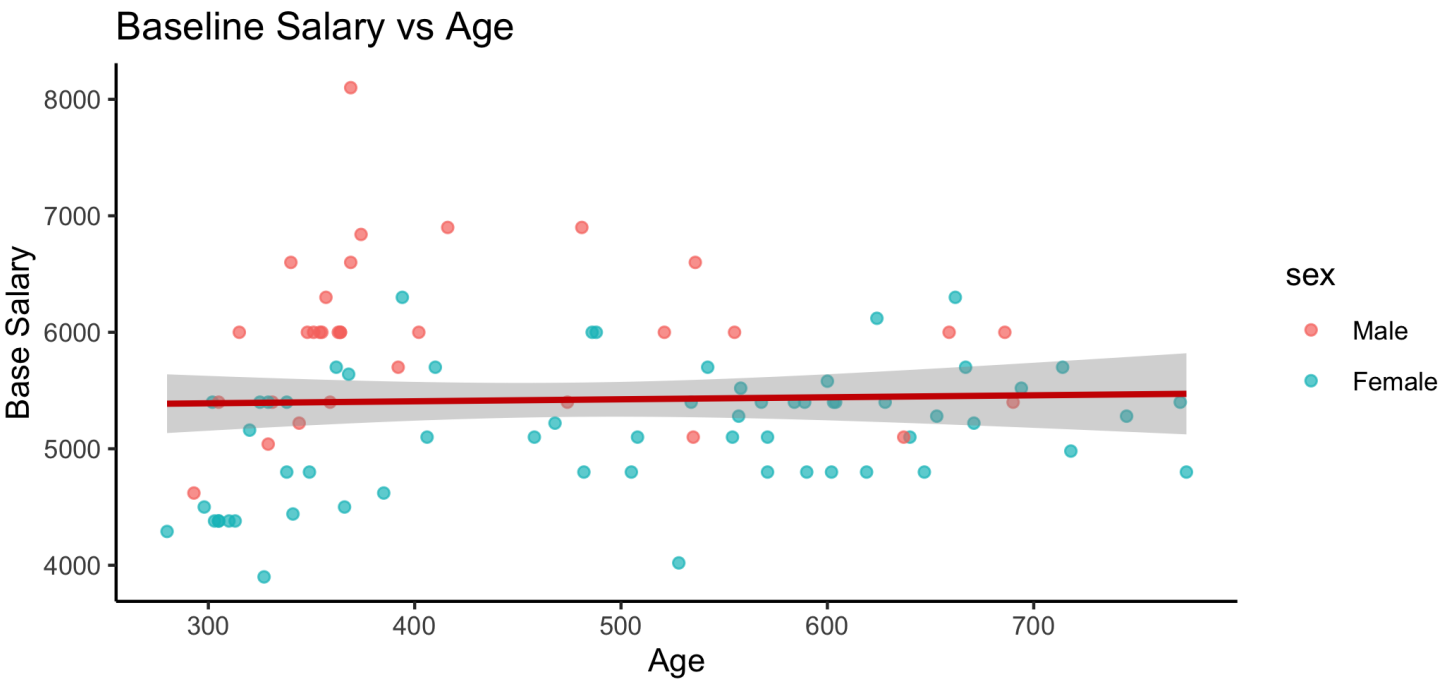
```
ggplot(wages,aes(x=senior, y=bsal)) +
  geom_point(alpha = .7,aes(color=sex)) +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Baseline Salary vs Seniority",x="Seniority",y="Base Salary")
```



# EDA

Next, bsal vs. age

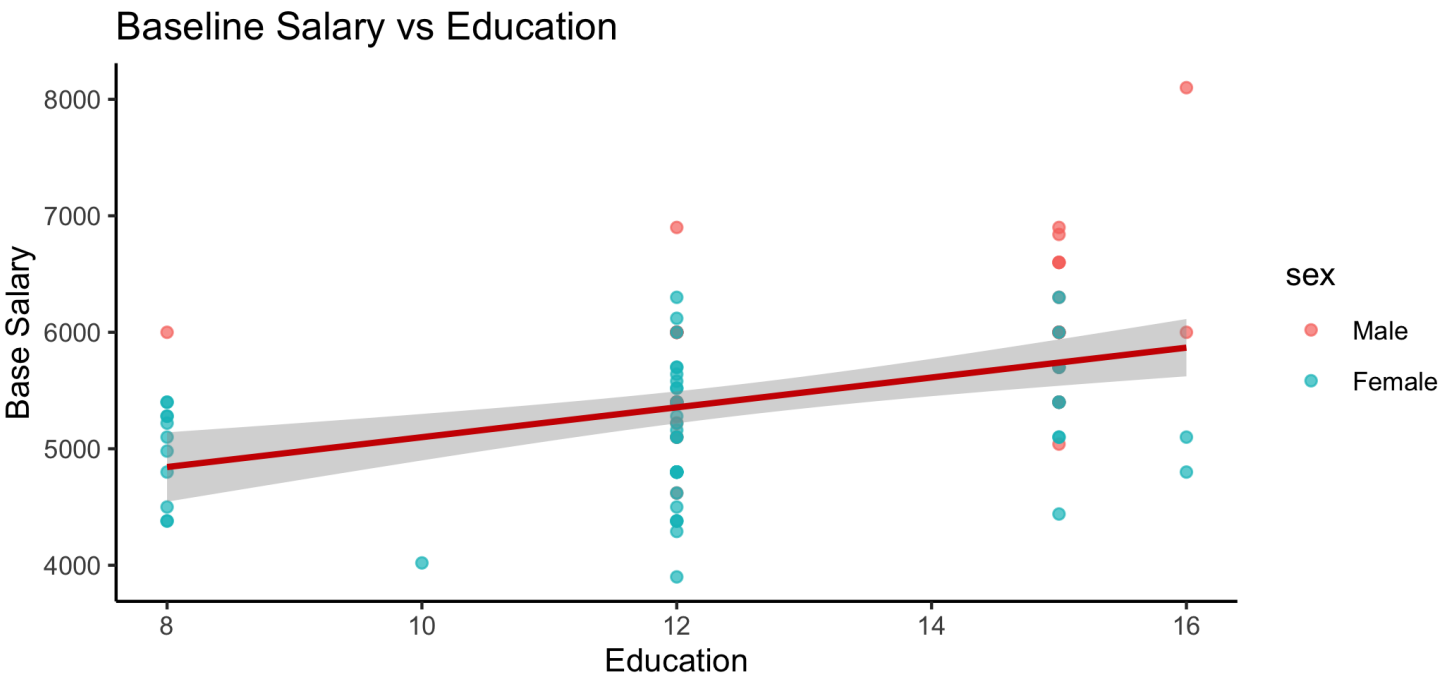
```
ggplot(wages,aes(x=age, y=bsal)) +
  geom_point(alpha = .7,aes(color=sex)) +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Baseline Salary vs Age",x="Age",y="Base Salary")
```



# EDA

## bsal vs. educ

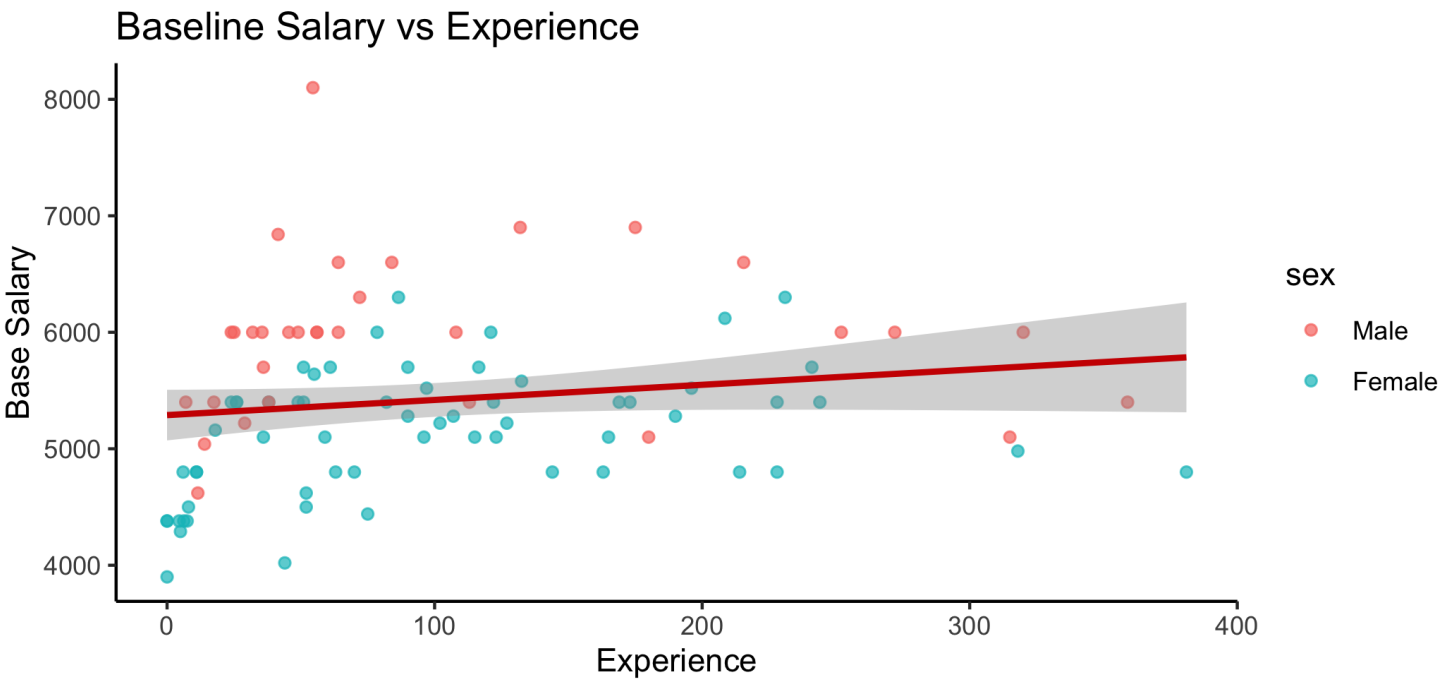
```
ggplot(wages,aes(x=educ, y=bsal)) +
  geom_point(alpha = .7,aes(color=sex)) +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Baseline Salary vs Education",x="Education",y="Base Salary")
```



# EDA

Finally, bsal vs. exper

```
ggplot(wages,aes(x=exper, y=bsal)) +
  geom_point(alpha = .7,aes(color=sex)) +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Baseline Salary vs Experience",x="Experience",y="Base Salary")
```



# TAKEAWAYS

- Clearly, they are other variables that may be relevant in explaining baseline salary.
- We need to explore other statistical methods than the t-test and simple linear regression.
- We need methods that can explore the relationship between baseline salary and sex while also controlling for the other variables that clearly may be relevant.
- This brings us to multiple linear regression (MLR).
- Something to keep in mind, the overall conclusions may not change after using a better model for this data.

  In general, this should never stop you from exploring and reporting the results from better models; you should always be rigorous when doing analyses and be honest when reporting the results!

# What's next?

Move on to the readings for the next module!