DUNMIES, FACTORS, ANOVA, AND LINEAR REGRESSION

Review

https://ai.lange-analytics.com/

WHAT WILL YOU LEARN

- Review: Dummy Variables in Linear Regression
- One-Way ANOVA
- AB Tests
- Two-Way ANOVA

LIBRARIES AND DATA

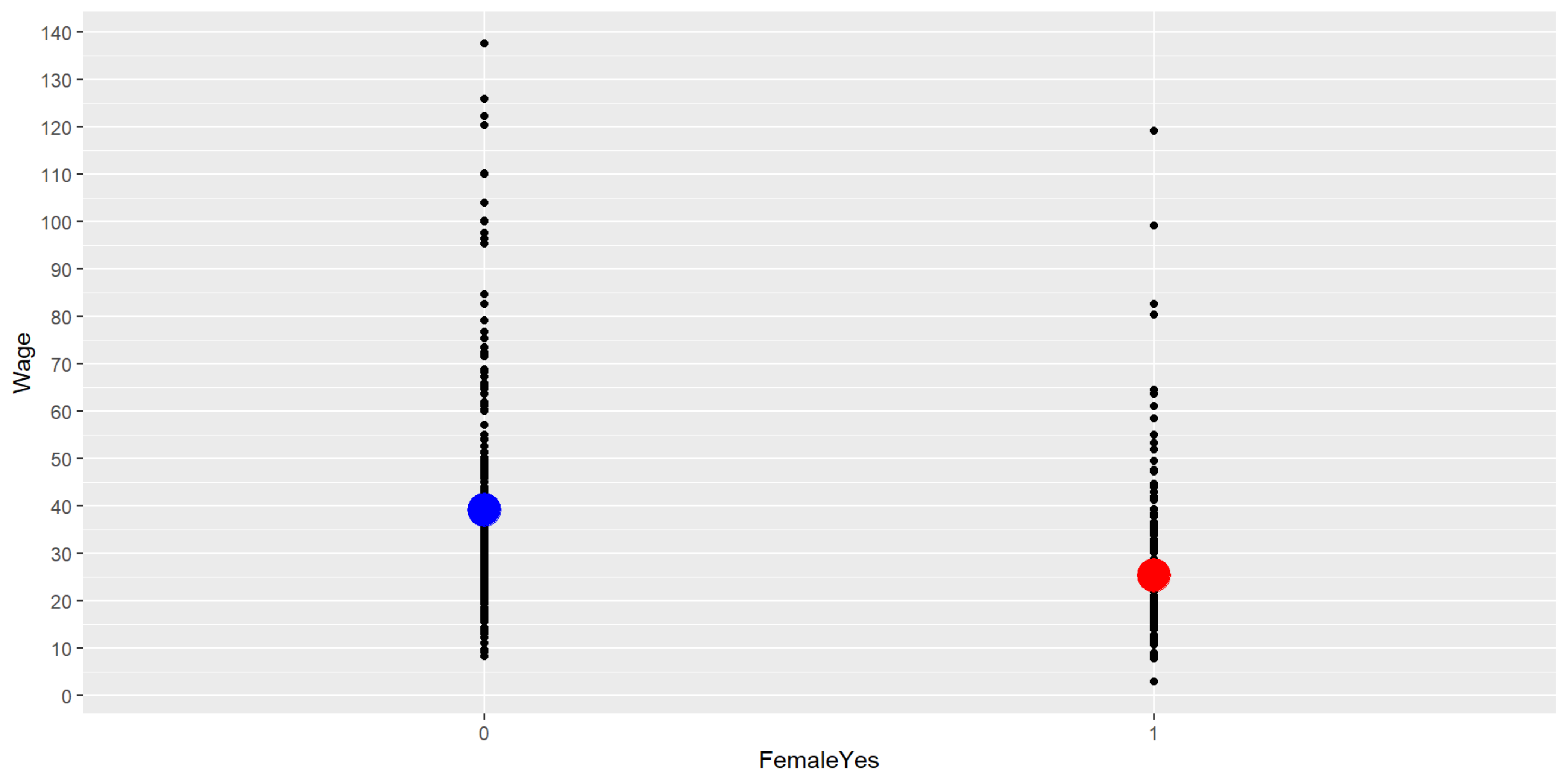
▶ Code

```
     Wage   Educ FemaleYes
1 17.0810   NoHS         1
2 17.8524     HS         1
3 16.5300   NoHS         0
4 33.0600   NoHS         0
5 29.2030     HS         0
6 48.2125 Degree         0
```

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

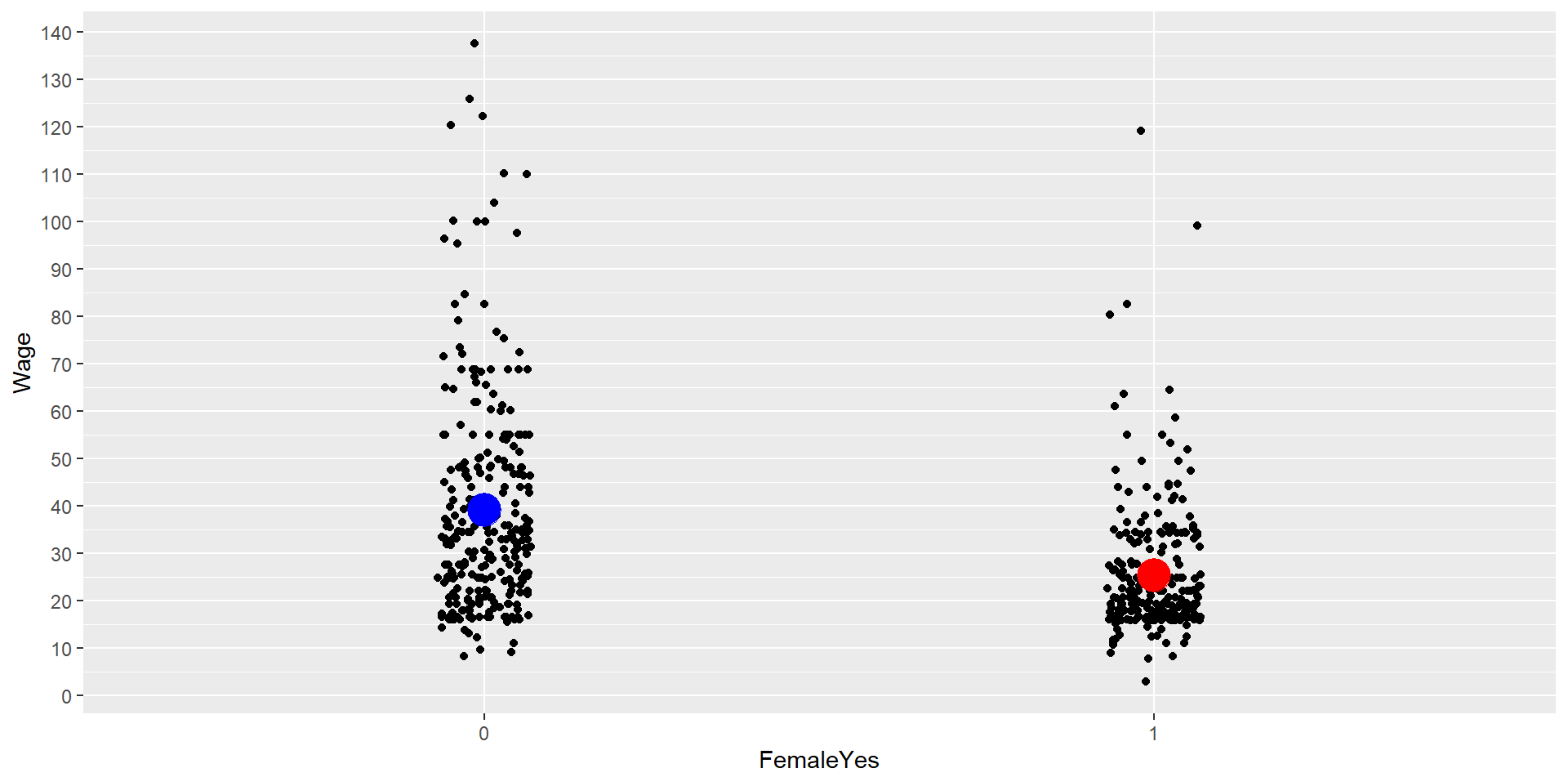
GRAPHICAL APPROACH (COMPARING MEANS)



https://ai.lange-analytics.com/

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

GRAPHICAL APPROACH (COMPARING MEANS)



https://ai.lange-analytics.com/

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

T-TEST (COMPARING MEANS)

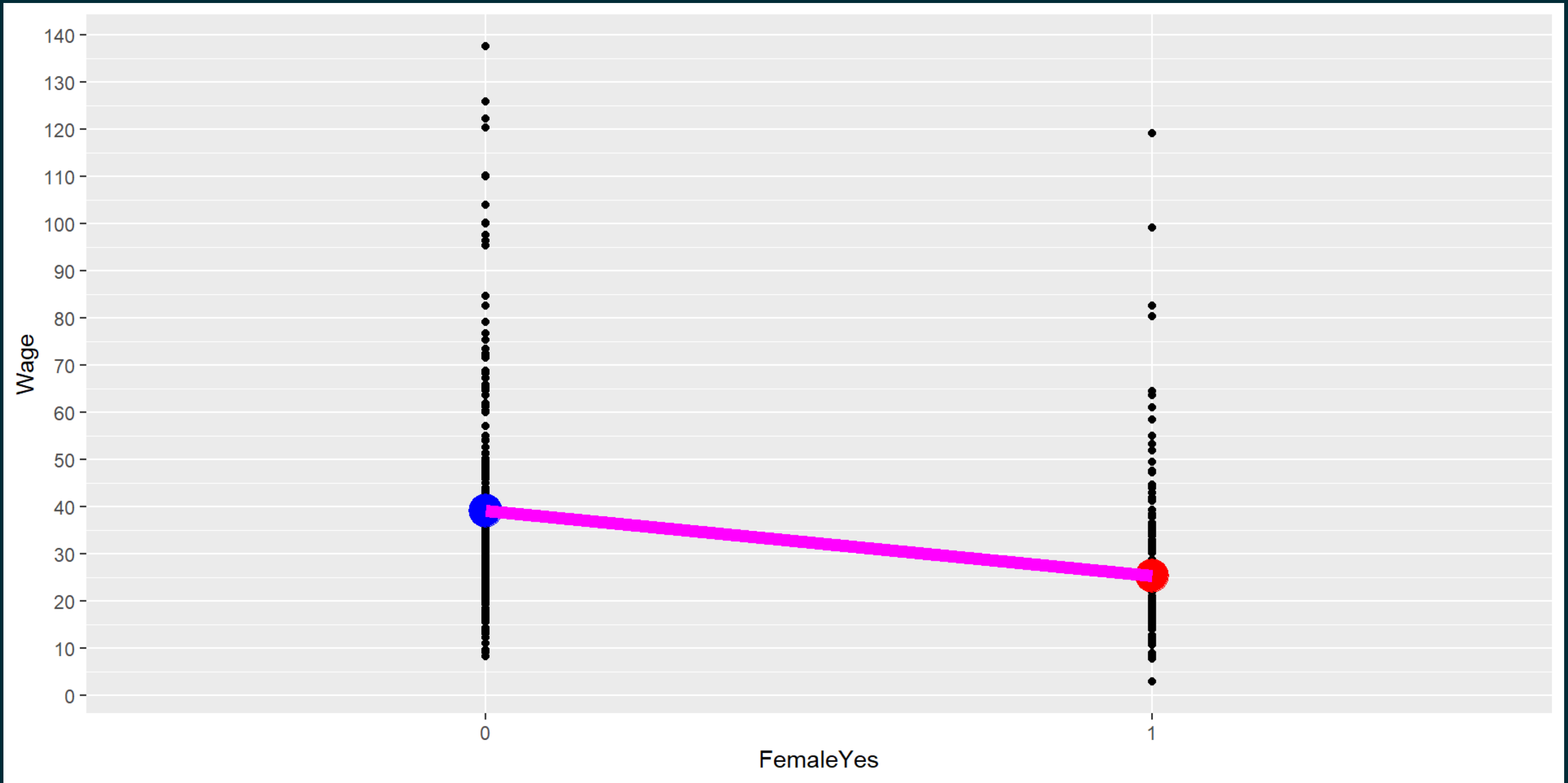
▶ Code

```

	Two Sample t-test

data:  DataWageFem$Wage and DataWageMale$Wage
t = -8.2199, df = 522, p-value = 1.621e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.09015 -10.49698
sample estimates:
mean of x mean of y 
 25.32462  39.11818 
```

The difference in means is: -13.7935635

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

GRAPHICAL APPROACH (OLS REGRESSION)



https://ai.lange-analytics.com/

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

OLS REGRESSION APPROACH

▶ Code

```
Call:
lm(formula = Wage ~ FemaleYes, data = DataWage)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.853 -10.191  -5.489   8.020  98.522 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   39.118      1.159   33.75  < 2e-16 ***
FemaleYes1   -13.794      1.678   -8.22 1.62e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.19 on 522 degrees of freedom
Multiple R-squared:  0.1146,	Adjusted R-squared:  0.1129 
F-statistic: 67.57 on 1 and 522 DF,  p-value: 1.621e-15
```

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

INTERPRETATION OF DUMMY VARIABLES

▶ Code

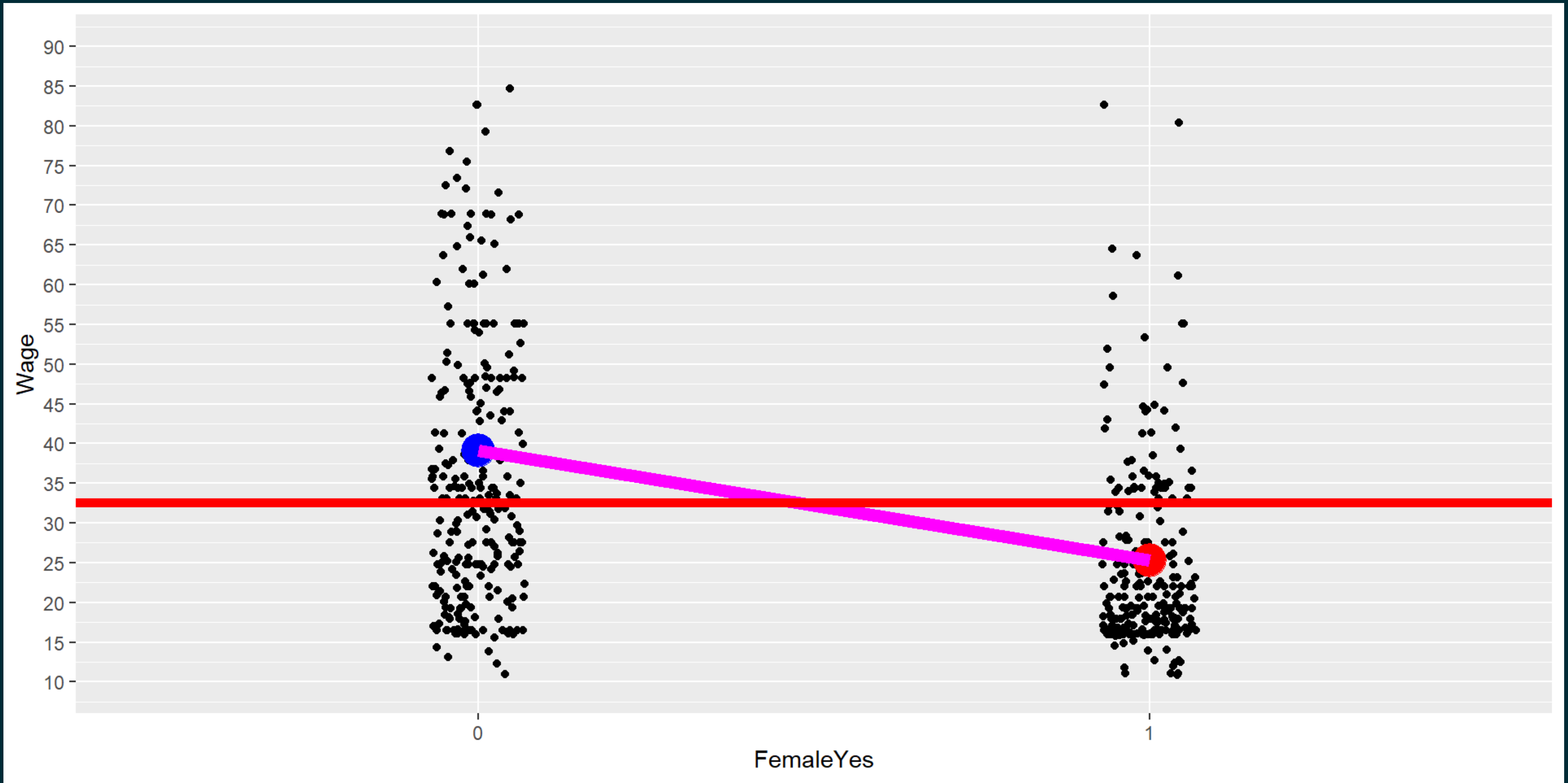
```
Call:
lm(formula = Wage ~ FemaleYes, data = DataWage)

Coefficients:
(Intercept)   FemaleYes1  
      39.12       -13.79  
```

$$\begin{aligned} Wage &= \beta_1 & \cdot FemaleYes & & + \beta_0 \\ Wage &= (-13.8) & \cdot FemaleYes & & + 39.1 \\ [-13.8] &= (-13.8) & \cdot [+1] & & + [+0] \end{aligned}$$

$FemaleYes$ can only increase by +1 from 0 to 1 — when a observation with $FemaleYes = 0$ (male) switches to $FemaleYes = 1$ (female). The consequence is that wage changes with -13.8.

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

ANOVA



https://ai.lange-analytics.com/

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

ANOVA

$$\text{Mean Total Error (Variance): } MTE \approx \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$$

$$\text{Mean Residual Error: } MSE \approx \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

F-Value:

$$F = \frac{MTE}{MSE} = \frac{\text{Mean Total Error (Restricted Model)}}{\text{Mean Residual Error (Full Model)}}$$

ARE WAGE MEANS DIFFERENT FOR FEMALES AND MALES?

ANOVA

$$F=\frac{MTE}{MSE}=\frac{\text{Mean Sum of Errors: Restricted Model}}{\text{Mean Sum of Errors: Full Model}}$$

▶ Code

```
Analysis of Variance Table

Response: Wage
           Df Sum Sq Mean Sq F value    Pr(>F)    
FemaleYes   1  24872 24872.1  67.567 1.621e-15 ***
Residuals 522 192153   368.1                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MTE= 24872.09 + 368.1086= 25240.2

If F>1. I.e., full model mean squared error is smaller than the one of the restricted model. And if this is not by chance (P is very small), then the variables from the full model must be significant.

ARE EDUCATION DIFFERENCES RELEVANT FOR WAGES?

DATA

▶ Code

```
     Wage   Educ FemaleYes
1 17.0810   NoHS         1
2 17.8524     HS         1
3 16.5300   NoHS         0
4 33.0600   NoHS         0
5 29.2030     HS         0
6 48.2125 Degree         0
```

▶ Code

```
[1] "NoHS"   "HS"     "Degree"
```

ARE EDUCATION DIFFERENCES RELEVANT FOR WAGES?

(PAIR WISE) T-TEST

Pair-wise t-test is problematic to indicate if factor is relevant because of multi-testing problem.

ARE EDUCATION DIFFERENCES RELEVANT FOR WAGES?

ONE-WAY ANOVA (ONE FACTOR (EDUCTION) WITH 3 GROUPS)

▶ Code

```
Call:
lm(formula = Wage ~ Educ, data = DataWage)

Residuals:
   Min     1Q Median     3Q    Max 
-29.71 -11.69  -5.69   6.93  96.91 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22.403      1.786  12.547  < 2e-16 ***
EducHS         7.193      2.241   3.209  0.00141 ** 
EducDegree    18.330      2.214   8.278 1.06e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.06 on 521 degrees of freedom
Multiple R-squared:  0.1275     Adjusted R-squared:  0.1241
```

▶ Code

```
Analysis of Variance Table

Response: Wage
           Df Sum Sq Mean Sq F value    Pr(>F)    
Educ        2  27661 13830.7  38.053 3.758e-16 ***
Residuals 521 189363   363.5                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MTE = 13830.7 + 363.4614 = 14194.16$$

ARE EDUCATION DIFFERENCES RELEVANT FOR WAGES?

ADJUSTED PAIRED T-TEST

▶ Code

```
	Pairwise comparisons using t tests with non-pooled SD 

data:  DataWage$Wage and DataWage$Educ 

       NoHS    HS     
HS     2.7e-05 -      
Degree < 2e-16 2.2e-07

P value adjustment method: bonferroni 
```

AB-TEST

A very powerful and easy to use methodology to compare means of one or more *Treatment* groups to a *Control Group*.

Goal: Determine if a treatment(s) (e.g., conversion rates for one or more new websites (*Treatment*)) are significant compared to the old website (*Control Group*).

Methodology: To evaluate significant differences between groups *One-Way ANOVA* possibly followed by a *post-hoc* pairwise t-test, if *ANOVA* was successful and the dependent variable is continuous. For binary dependent variables alternatives to ANOVA exist.For experiments with binary outcome and one control group and one treatment group a t-test can be used.

Requirement: Participants must be randomly assigned to the groups (no self-selection!).

Problem: While it is easy and technically straightforward to randomly assign website visitors to different webpages, it can be difficult or impossible in other cases.

For example, it is not feasible to assign participants of a marketing event into a treatment group (drank a glass of champagne before the talk) and a control group (did not drink a glass of champagne before the talk)

TWO-WAY ANOVA (NOT RELEVANT FOR MIDTERM)

WITH INTERACTION TERM

▶ Code

```
Analysis of Variance Table

Response: Wage
                Df Sum Sq Mean Sq F value    Pr(>F)    
Educ             2  27661 13830.7 43.0867 < 2.2e-16 ***
FemaleYes        1  21998 21997.6 68.5291 1.069e-15 ***
Educ:FemaleYes   2   1089   544.6  1.6966    0.1843    
Residuals      518 166277   321.0                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TWO-WAY ANOVA (NOT RELEVANT FOR MIDTERM)

WITH INTERACTION TERM (AGAIN WITH DIFFERENT ORDER)

▶ Code

```
Analysis of Variance Table

Response: Wage
                Df Sum Sq Mean Sq F value    Pr(>F)    
FemaleYes        1  24872 24872.1 77.4838 < 2.2e-16 ***
Educ             2  24787 12393.5 38.6093 2.345e-16 ***
FemaleYes:Educ   2   1089   544.6  1.6966    0.1843    
Residuals      518 166277   321.0                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TWO-WAY ANOVA (NOT RELEVANT FOR MIDTERM)

WITHOUT INTERACTION TERM

▶ Code

```
Analysis of Variance Table

Response: Wage
           Df Sum Sq Mean Sq F value    Pr(>F)    
FemaleYes   1  24872 24872.1  77.277 < 2.2e-16 ***
Educ        2  24787 12393.5  38.506 2.542e-16 ***
Residuals 520 167366   321.9                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TUKEY PAIRWISE T-TEST (NOT RELEVANT FOR MIDTERM)

WITHOUT INTERACTION TERM

▶ Code

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Model2WayWageNoInt)

$FemaleYes
         diff       lwr     upr p adj
1-0 -13.79356 -16.87613 -10.711     0

$Educ
                 diff       lwr      upr    p adj
HS-NoHS      9.015023  4.057401 13.97264 6.77e-05
Degree-NoHS 17.940746 13.043300 22.83819 0.00e+00
Degree-HS    8.925723  4.758252 13.09319 2.00e-06
```

https://ai.lange-analytics.com/

https://ai.lange-analytics.com/