E/M 150: Applied Business Analytics Activity 4
Due Dates

- • Individual submission: 8am on Thursday November 3rd
- • Group submission: 11:59 pm on Friday November 4th

Background: Working on an assembly line often involves performing tedious and repetitive tasks leading to high worker turnover. The production manager at a factory that uses assembly line production wants to develop a model to predict whether a newly hired worker will stay on the job for at least one year. Employment records for 32 employees are chosen at random. For each employee the following information is obtained:

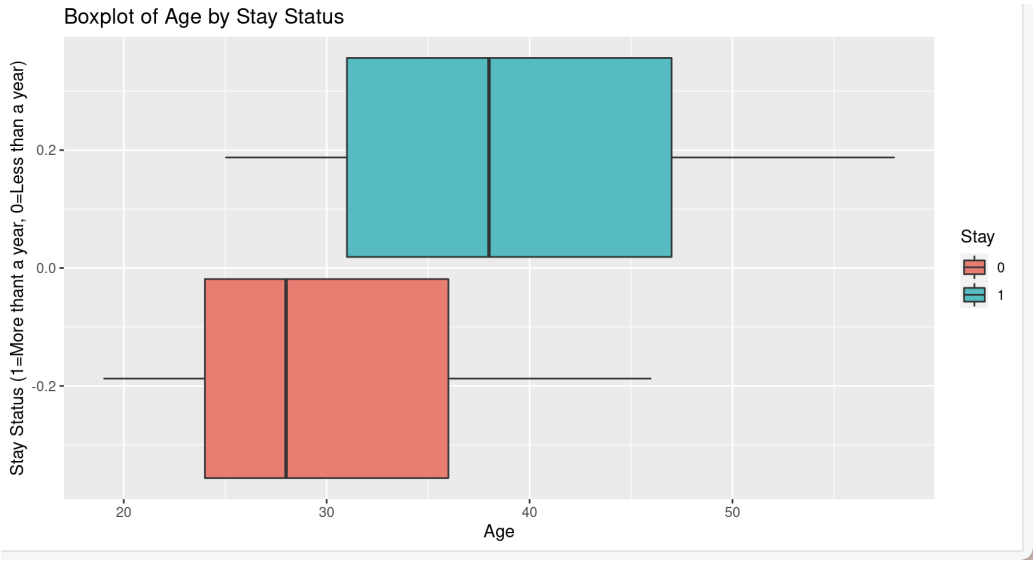
- • Stay: whether the employee stays for at least one year (1=yes, 0=no)
- • Age: age of employee at time of hire in years
- • Female: whether the employee identifies as female (1=yes, 0=otherwise)
- • Assembly: whether the employee has worked on an assembly line before (1=yes, 0=no)

1. First check variable type and change type as needed. [1pt]
   - All variables are numerical
   - Changed stay, assembly and female from numerical to categorical variables
2. What is the response variable for this analysis? Identify if the variable is categorical or numeric. [1pt]
   - The response variable is stay, whether the employee stays for at least one year or not
   - The response variable is categorical
3. Let's first look at the relationship between whether an employee stays for at least one year and age.

   a. Obtain the median age and IQR of age for workers that stay for at least a year and those that do not (perhaps use the handy table below).[2pts]

	Stay for at least a year	Did not stay for at least a year
Median	38	28
IQR	16	12

   b. Obtain the appropriate graphical display to examine how the distribution of age differs by stay status. [2pts]



c. Using the above summary statistics and plot, compare the workers that do and do not stay for at least a year in terms of age. [3pts]

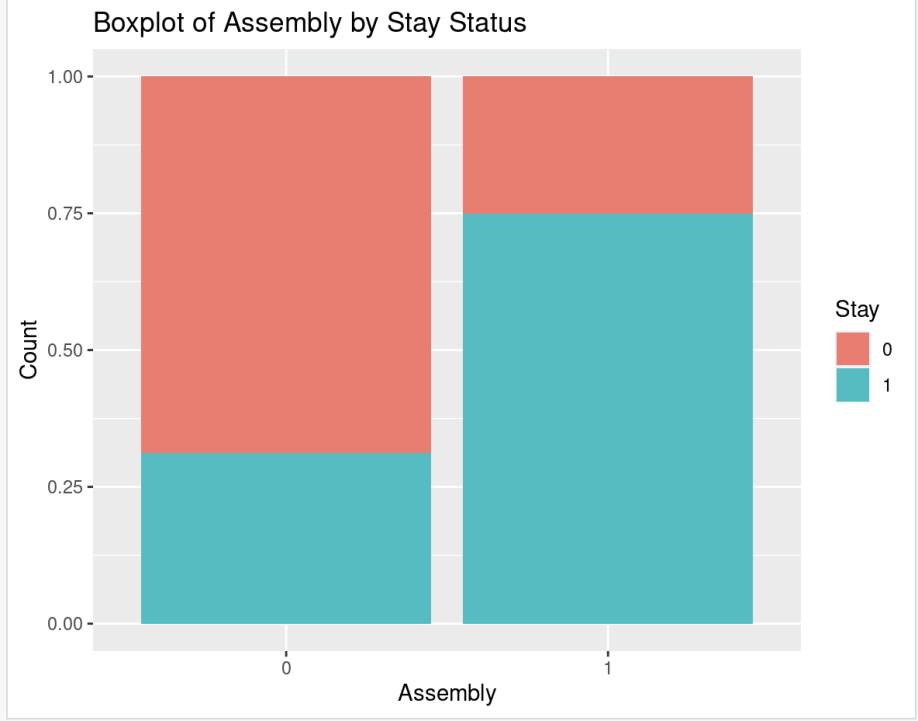
- When comparing the distribution of stay status by age, the most noticeable difference is that the median for people that stayed for at least a year is much higher than the median for people that did not stay for a year (38 vs 28). The shape is slightly skewed right for both Stay status and the shape is larger for those who stay for at least a year. There are no potential outliers for each group.

4. Next, let's look at the relationship between whether an employee stays for at least one year and whether they have worked on an assembly line before.

a) Construct (and provide) the appropriate table that examines whether the proportion of workers that stay for at least one year differs by previous assembly line experience. Remember, do NOT copy and paste console output from RStudio! [2pts]

	Assembly (0, No)	Assembly (1, Yes)	N/A
Stay (0, No)	68.75	31.25	
Stay (1, Yes)	25	75	
N/A			

b. Obtain the appropriate graphical display that examines whether the proportion of workers that stay for at least one year differs by assembly line experience. [2pts]



c. Using your table and plot from above, explain if there is a relationship between previous assembly line experience and whether a worker stays for at least one year. [3pts]

The percentage of people that stay that have worked on an assembly line before is approximately 55 percentage points higher than the percentage of people that stay but have not worked on an assembly line (31.25% vs 75%). This is a very noticeable and quite large difference.

5. Recall that with logistic regression we typically describe results in terms of odds. Calculate and interpret the odds of staying for at least one year. [2pts]

- For every person that does not stay for over a year, 1.13 stay for over a year.
- In better terms, for every 10,000 people that do not stay for over a year, 11300 stay for over a year.

6. Fit the logistic regression model where we are predicting whether an employee stays for at least one year based on Age, Female, and Assembly.

a. Report the estimated logistic regression model. [2pts]

- log(Phat/1-phat)=-4.20+0.09Age+0.08Female+1.78Assembly

b. Interpret the slope coefficient for Age by filling in the following blanks. [2pts]

For each additional year of age, it is estimated that the odds of staying for at least one year increase by 9.4% holding the assembly and female variables constant

c. Interpret the slope coefficient associated with the variable Assembly by filling in the following blanks. [2pts]

It is estimated that odds of workers with prior assembly line experience staying for at least a year are 492% higher than the odds of workers without prior assembly line experience staying for at least a year holding the age and female variables constant.

7. It is believed that older workers are more likely to stay for at least a year. Conduct the associated hypothesis test to evaluate this belief by completing the following parts. a. Set up the null and alternative hypotheses. [1pt]

H0: $\beta_{age}=0$

HA: $\beta_{age}>0$

b. Provide the associated test statistic. [0.5]

- 1.689 = z stat

c. Provide the associated p-value [0.5]

- 0.0456

d. Provide your decision and a justification for this decision. [1pt]

- Reject the null because 0.0456<0.05

e. Provide a conclusion for the hypothesis test. [2pts]

- At the 5% significance level we reject the null hypothesis and conclude that there is very strong evidence that the odds of Older workers staying longer than a year is higher than the odds of a younger person staying at least a year holding gender and assembly experience constant.
- Z stat= 1.689, p value= 0.0456

8. Next let's look at the confidence interval associated with females.

a. Report and interpret the 95% confidence interval for the slope coefficient for Female. Remember that you will need to backtransform! [3pts]

Log odds : (-0.8783, 2.6226)

Odds:(-58.54, 1273.572)

We are 95% confident that the odds of workers identifying as female stay for at least a year is between 58.54% lower to 1273.572% higher than the odds of a worker that does not identify as a female.

b. Based on the confidence interval, explain if there is evidence to suggest that the odds of workers identifying as female staying for at least a year differ from the odds of those not identifying as female staying for at least one year. [2pts]

Since 0 is included in the 95% confidence interval, there isn't very strong evidence to suggest that the odds of workers identifying as female stay for at least a year differ from those that do not identify as a female.

9. Sofia is a 40-year-old worker who identifies as female and has no previous assembly line experience. What is the probability that Sofia stays for at least one year? [2pts]

- 54.94%

10. In class we discussed four different assumptions that must be reasonably satisfied for the results of our logistic regression to be valid. For each assumption, provide a brief description/explanation of the assumption (i.e., what are we assuming with the independence assumption) and whether the assumption is reasonably satisfied or not and why. [4pts]

Assumption	Met or Not	Why
Independence	Met	There was a random sample
Binary Response	Met	The response variable has two categories, stayed for a year of not
Linearity	Assume Met	There is no obvious reason to doubt it
Large Sample Size	Not Met	The sample size was less that 500 (32)