Q1

What is the attrition rate for employees at BI? (A rate, remember, is expressed as a proportion.)

1. Calculate overall attrition rate.
2. Create a summary table of conditional attrition rates by department and job role. (The table should have 3 columns: department, job role, and the calculated conditional attrition rate.) Sort this table by attrition rate in descending order.

Note: The simplest possible classification model would be to use the attrition majority class—"Yes" or "No"—as the prediction. This is called "majority class" prediction. The in-sample accuracy of the majority class model is simply the proportion of the majority class. This is an important performance benchmark.

Q2

Fit a logistic regression model of attrition using all the predictors. (Note: employee_number is NOT a predictor!)

1. Report accuracy for this model with a decision threshold of .5. (Accuracy is defined as the proportion of correct predictions.)
2. Comment on whether the model offers an improvement over predicting with the majority class.

Q3

The upside of standardizing inputs by centering and scaling is that it allows you to compare coefficient effect sizes easily—they are all on the same scale. (The downside is that they are no longer scaled in the original units, and interpretation changes.) Even though the coefficients are expressed in log odds in this case, after standardization they can still be compared for effect sizes on a relative basis.

There are a lot of coefficients to type into the model formula. A shortcut to automatically include all the predictors in the dataset is `.`, as in: `glm(target ~ ., family = binomial, data = ...)`. However, this shortcut doesn't allow you to standardize also. The easiest solution to create a new data set in which all the continuous variables are centered. For this a version of `mutate()` is useful: `mutate_if()`. The code would go like this:

```
data %>%    mutate_if(is.numeric, scale)
```

In English: if the variable is numeric, then scale it.

Notice that some of the standard errors and coefficients in the model above have exploded. (You can see this more easily if you adjust the number of digits printed in the output with `options(scipen = 3)`.) The SEs for some of the `department` and `job_role` coefficients are over 380. Why has this happened? Multicolinearity! Some of the levels of the `department` variable are correlated with levels in `job_role`. For example, since most of the people in the Human Resources department also have a job title of Human Resources, the information from `department` is redundant: by definition, if we know `job_role` we also know `department` and vice versa. This is a textbook example of how multicollinearity makes inference difficult—we can't compare the coefficients because some of them are wacky. The solution? Remove the redundant variable. Refit the model without `department`

1. Which of the centered and scaled predictors has the largest effect size?
2. Interpret the coefficient with the largest effect size. Since you are working with standardized coefficients, the interpretation for continuous predictors will be: a 1 unit (that is, after scaling, a 1 standard deviation) increase in `x` is associated with a coefficient-sized change in the log odds of `y`, on average, while holding the other predictors constant. The coefficient represents the change in the log odds of the outcome associated with an increase from the reference level in the categorical variable.

Q4

Based on the above logistic regression model (and, specifically, on the coefficient with the largest effect size that you identified above), **how might company policy be changed to reduce employee attrition**?

1. Describe your proposed policy change.

2. Estimate and explain the change in churn probability associated with that policy change.

Q5

What should Angelica say in her report? Please include quantitative details from your answers to the questions above.

Challenge Question

Write a practice question for the Logisitc Regression section of *Learn R + Statistics*. It should explore the content in one of the lessons and be modeled on this question, with both a hint and an answer.