Group Work 08

```
library(bayesrules) # R package for our textbook
library(tidyverse) # Collection of packages for tidying and plotting data
library(janitor) # Helper functions like tabyl
library(rstan) # for fitting models
library(rstanarm) # for fitting standard regression models
library(broom.mixed) # for tidy() function
library(bayesplot) # helpful plotting functions
library(tidybayes) # helpful for wrangling Bayesian model output
library(patchwork)
```

1 BR 13.2

Calculate and interpret the odds for each event of interest below.

(a) The probability of rain tomorrow is 0.8.
(b) The probability of flipping 2 Heads in a row is 0.25.
(c) The log(odds) that your bus will be on time are 0.
(d) The log(odds) that a person is left-handed are -1.386.

2 BR 13.14 (adapted)

We return to the `pulse_of_the_nation` survey data in the {bayesrules} package which includes a variable on whether or not a person believes in `ghosts` (spooky!) You are going to build a logistic regression model using `age`, `education`, and `books` as predictors.

(a) Perform an appropriate EDA of the relationship between `ghosts` and each of the predictors
(b) Specify the full Bayesian model
(c) Fit the model specified in (b) using `stan_glm`. Include MCMC convergence diagnostics.
(d) Report the `tidy` coefficient table with 90% credible intervals and interpret the coefficients in context

(e) Check the fit of the model using an appropriate posterior predictive check

3 BR 13.5

The confusion matrix below summarizes the performance of a logistic model in classifying the beliefs of 1000 survey respondents, using a probability cut-off of 0.5.

	y=0	y=1
predict FALSE	50	300
predict TRUE	30	620

(a) Calculate and interpret the model's overall accuracy.
(b) Calculate and interpret the model's sensitivity.
(c) Calculate and interpret the model's specificity.
(d) Suppose that researchers want to improve their ability to identify people that do *not* believe in climate change. How might they adjust their probability cut-off: Increase it or decrease it? Why?

4 BR 14.8/14.9 (adapted)

The `fake_news` data in the {bayesrules} package contains information about 150 news articles, some real news and some fake news. In the next exercises, our goal will be to develop a model that helps us classify an article's type, real or fake, using whether an article's title has an exclamation point (`title_has_excl`), the number of words in the title (`title_words`), and the percent of words in the article that have a negative sentiment (`negative`).

(a) In using naive Bayes classification to classify an article's type based on its title_words, we assume that the number of title_words are conditionally Normal. Do you think this is a fair assumption in this analysis?
(b) Suppose a new article is posted online and its title has 15 words. Utilize naive Bayes classification to calculate the posterior probability that the article is real. Do so from scratch, without using `naiveBayes()` with `predict()`.
(c) Check your work to part c using `naiveBayes()` with `predict()`.
(d) Construct a cross-validated confusion matrix for the naive bayes model
(e) Starting from weakly informative priors, use `stan_glm()` to simulate the posterior logistic regression model of news type by all three predictors: `title_words`, `negative`, and `title_has_excl`. You should confirm the model is adequete with MCMC diagnostics and a `pp_check`.
(f) Using a probability cutoff of 0.5, obtain cross-validated estimates of the sensitivity and specificity for the logistic regression model.

(g) Compare the cross-validated metrics of the logistic regression model to those of naive_model_4. Which model is better at detecting fake news?
(h) If our goal is to best detect when an article is fake, which of the four models should we use?

5 Wed

6 Wed