# Name:

# Lab 2 (Chapters 1-6 Review): **HappyPlanetIndex** + **drinks**

Originally these data were collected by two separate research groups. One group was a think tank (New Economic Foundation) interested in measuring the happiness of different countries (they compiled the data seen in the data frame **HappyPlanetIndex**). The other was the World Health Organization interested in how much and what kind of alcohol different countries drank for the purpose of understanding health (the data in **drinks**).

To get more information about each data frame, check out these links:

* <https://www.rdocumentation.org/packages/Lock5withR/versions/1.2.2/topics/HappyPlanetIndex>
* <https://www.rdocumentation.org/packages/fivethirtyeight/versions/0.4.0/topics/drinks>

Here I’ve provided the code to join these two data frames together so that each country’s happiness is aligned with their drinking data.

drinks$Country <- drinks$country

happydrinks <- left\_join(HappyPlanetIndex, drinks, by = "Country")

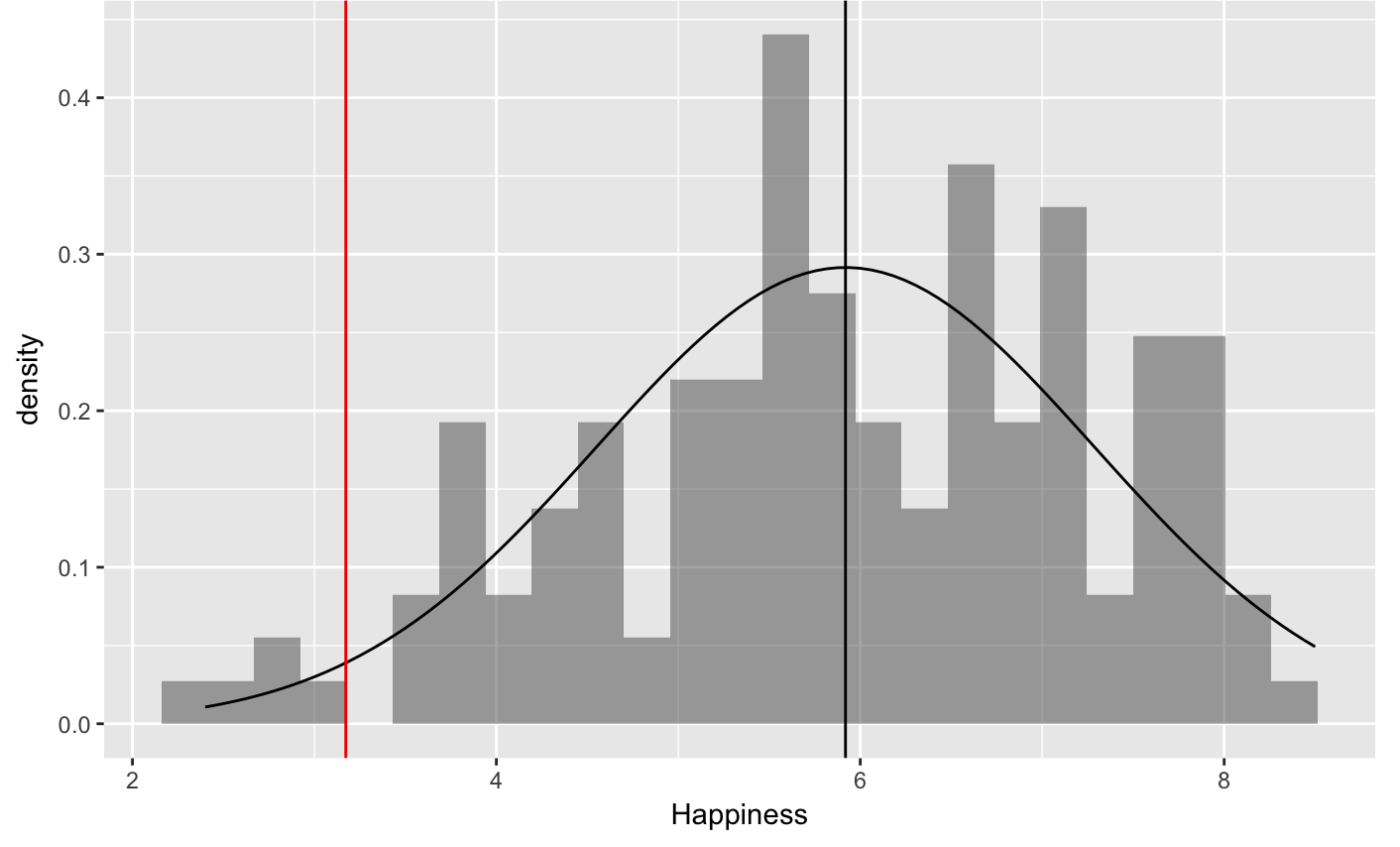
This code will create a new data frame called **happydrinks**: a dataset with 143 observations on the following 11 variables.

* Country Name of country
* Region Region in the world
* Happiness Score on a 0-10 scale for average level of happiness (10 is happiest)
* LifeExpectancy Average life expectancy (in years)
* Footprint Ecological footprint -- a measure of the (per capita) ecological impact
* HLY Happy Life Years - combines life expectancy with well-being
* HPI Happy Planet Index (0-100 scale)
* HPIRank HPI rank for the country
* GDPperCapita Gross Domestic Product (per capita)
* HDI Human Development Index
* Population Population (in millions)
* country name of country
* beer\_servings Servings of beer in average serving sizes per person
* spirit\_servings  Servings of spirits in average serving sizes per person
* wine\_servings  Servings of wine in average serving sizes per person
* total\_litres\_of\_pure\_alcohol  Total litres of pure alcohol per person

1. Countries vary in happiness. Which of these variables might help you predict how happy a country is? Come up with some ideas by looking at the variables in this data set.
2. How might someone use this data to explore whether alcohol consumption explains variation in happiness? Give an example of a pattern that would support their idea. (Note: this pattern does not have to be true in the data)

1. Which of the variables in the new data frame **happydrinks** are quantitative variables?

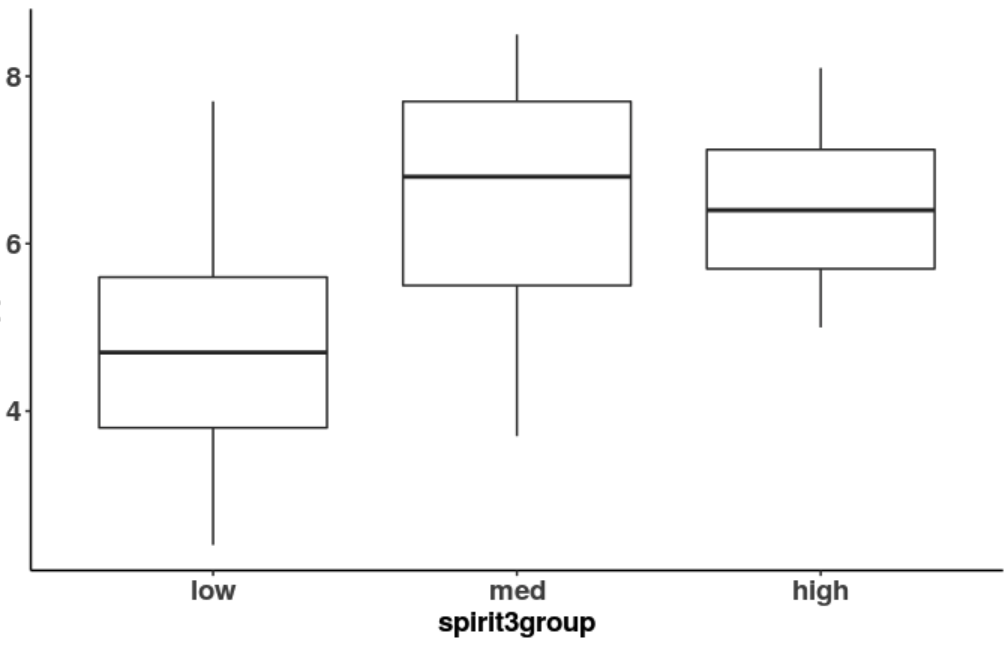
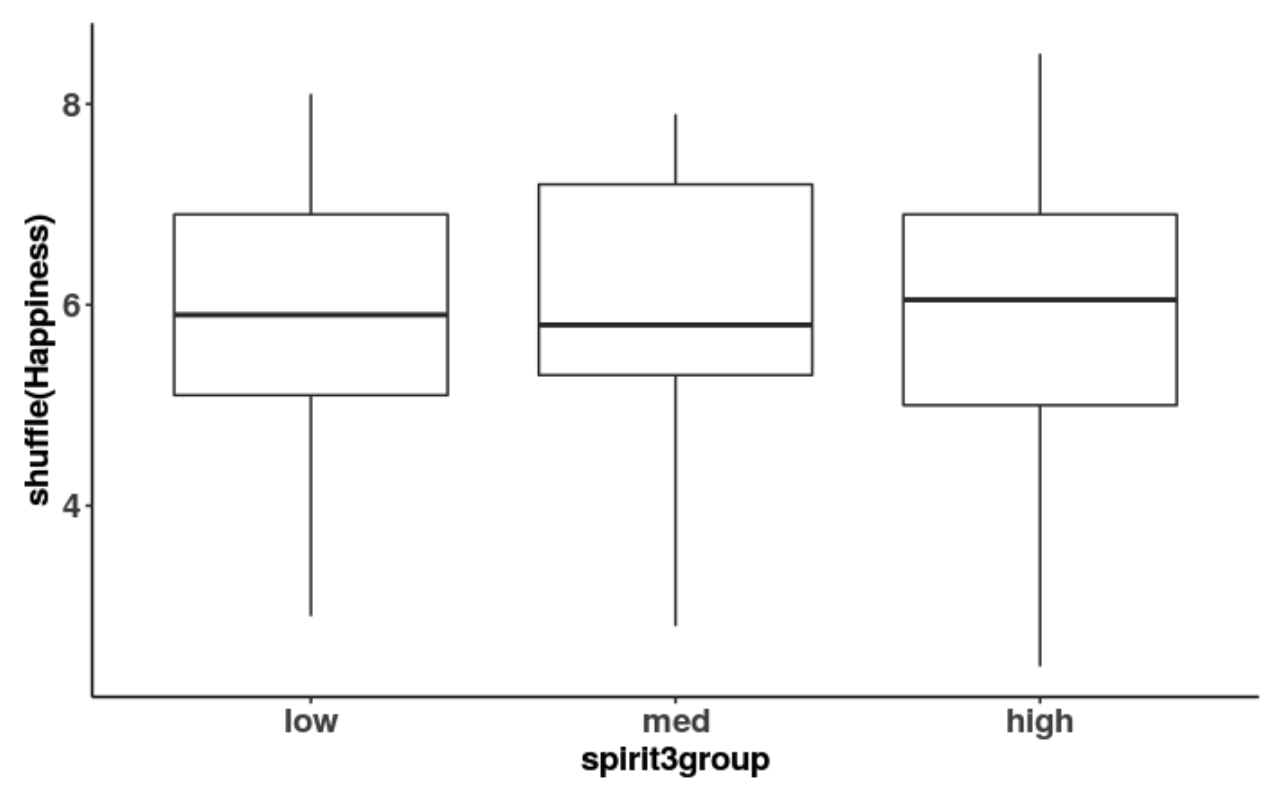
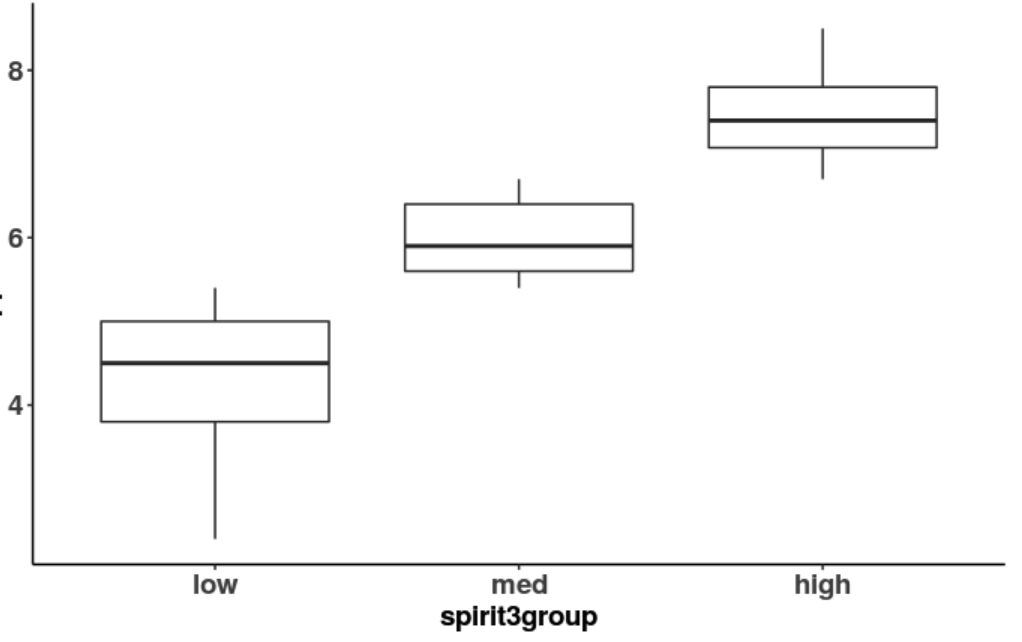
1. What are the cases in this data frame?
2. This data frame is terribly labeled. Label **Region** appropriately using R code. The regions are numbered from 1 to 7 and correspond to these labels. Write the R code below.
   * + 1. Latin America
       2. Western Nations
       3. Middle East and North Africa
       4. Sub-Saharan Africa
       5. South Asia
       6. East Asia
       7. Former Communist Countries
3. Write R code to create a histogram that shows you the variation in **Happiness**.
4. If you created a histogram of **Happiness** but added bins = 10, you would get a different looking histogram. Why is that?
5. How happy are you today? Rate on a scale of 1 - 10 (10 being very happy). Can we put this data into this data frame as a new row? Why or why not?
6. If you use the sample distribution of **Happiness** in the **happydrinks** data frame as a probability model, what is the likelihood that a country will have a happiness rating of 3 or lower?
7. In the figure below, which part represents the probability that a country would have a **Happiness** score of 3 or lower (depicted in a lighter color)? Which part represents the z score for a happiness rating of 3?



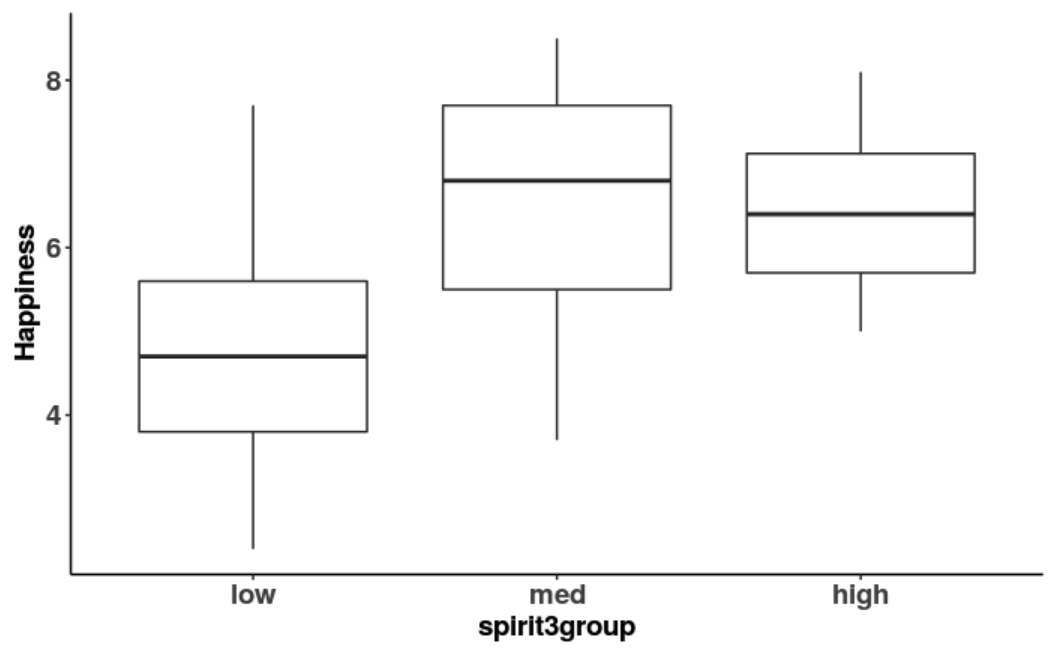
1. Using the normal distribution to model error (as in the picture above), what is the likelihood that a country will have a happiness rating of 3 or lower? Choose an answer below.
   1. Around .60
   2. Around .50
   3. Around .20
   4. Less than .10
2. These data are happiness ratings measured around 2009-10. If we measured the happiness of countries again, would you expect the exact same distribution of happiness? What would be different? What might be similar? (This question is about thinking/arguing; not a narrow right or wrong answer.)
3. Find the z-scores for all the countries’ **Happiness**. What’s Guatemala’s z-score? What does that number mean? What is Bangledesh’s z-score? What does that number mean?
4. Without looking at the data, write a story of the DGP arguing that the consumption of spirits would help explain the variation in happiness. Also, write your story of the DGP as a word equation.
5. Create at least two different visualizations to see whether **spirits3group** explains any of the variation in **Happiness** ratings. Write the code here.
6. Based on the visualizations, do you think **spirits\_servings** explains some of the variation in **happiness**? What do you see in the visualizations that makes you think that you could do a better job of predicting happiness if you knew about spirits drinking?
7. It almost seems like there are three types of spirit drinking countries -- those that don’t drink any, those that drink a medium amount of spirits, and those that drink a lot of spirits. Create a new categorical variable (call it **spirits3group**) in **happydrinks** that splits the countries up into these three groups. Label the groups as well. Write the R code here.
8. There are 9 countries that do not have a value for spirit\_servings (it’s NA). Write R code to put only the countries that have a value for **spirit\_servings** into a new data frame called **happy** data frame.

For the following questions, we will be using the **happy** data frame.

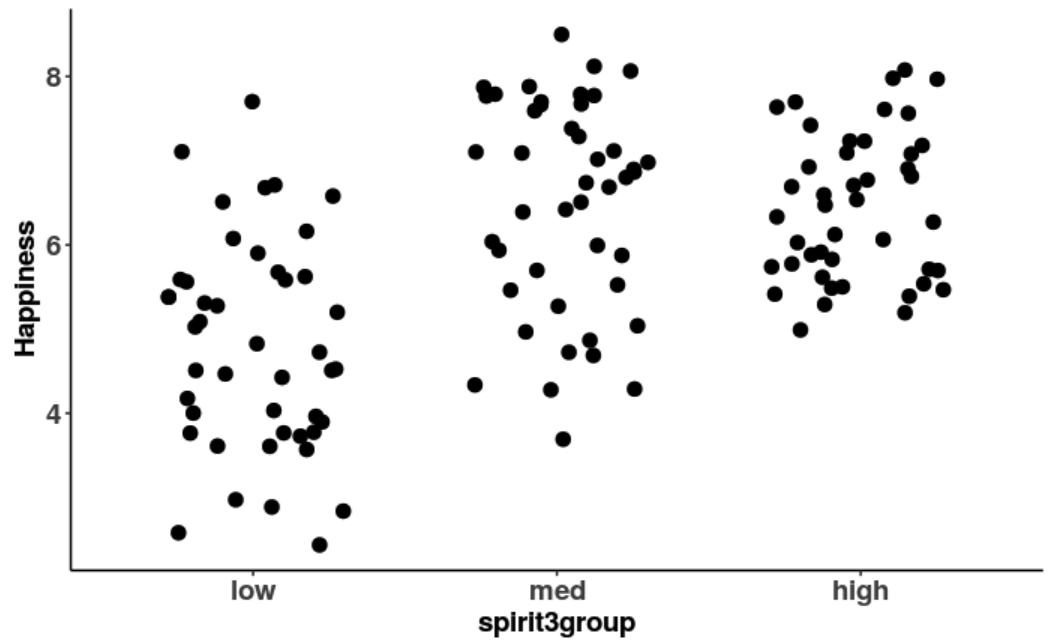
1. Make a histogram of **Happiness** by **spirit3group**. Put density plots over the histograms. Where on the density histogram would you look to see evidence of within group variation in **Happiness**?
2. Where on the density histogram would you look to see evidence of between group variation in **Happiness**?
3. Make a boxplot of **Happiness** by **spirit3group**. Where on the boxplot would you look to see evidence of within group variation in **Happiness**?
4. Where on the boxplot would you look to see evidence of between group variation in **Happiness**?
5. In our explorations so far, what is the outcome variable and what is the explanatory variable?
6. Based on our explorations of data, do you think drinking spirits *causes* countries to be more happy? Why or why not?
7. If we shuffled the **Happiness** scores of these countries and re-made our boxplot of **Happiness** by **spirit3group**, which pattern is more likely (choose one of the three pictures)? Why?

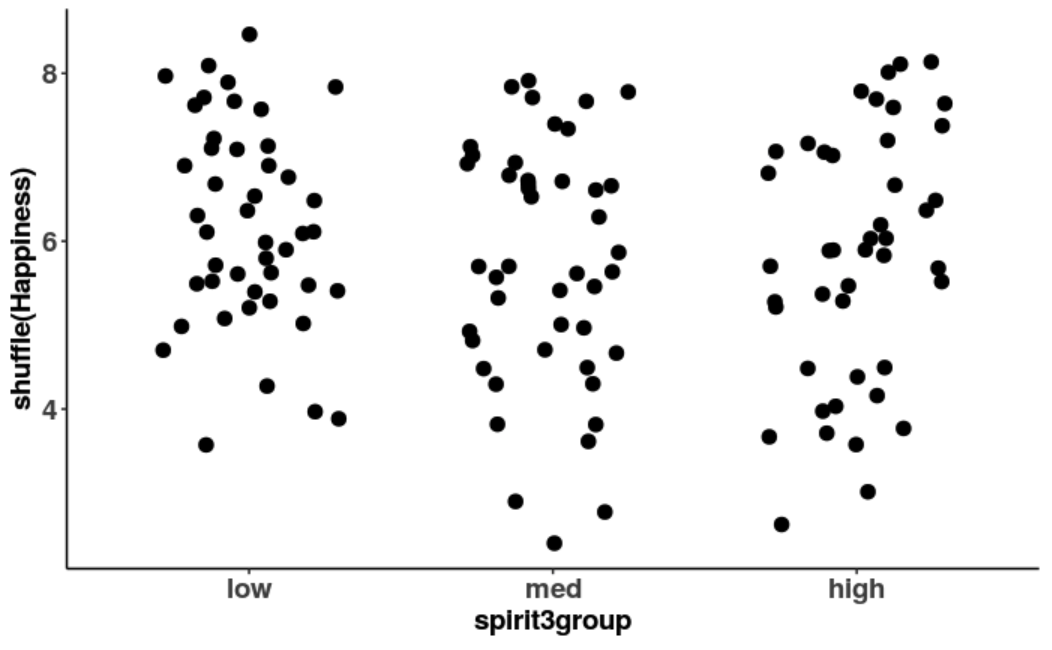
1. If we put plotted the empty model’s prediction on our boxplot (of our real data; not our shuffled data), where would those predictions go? Draw it in the boxplot provided here. What R code would help you put the predictions on the boxplot?



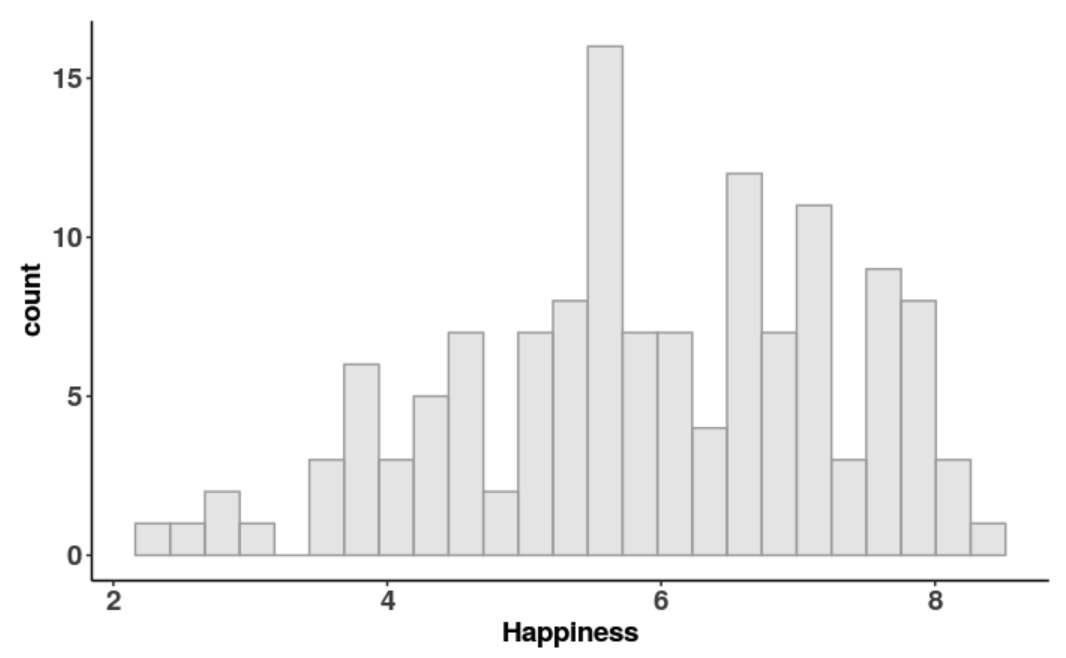
1. If we used the empty model to predict the happiness of some country, what value would we use?
2. Admittedly, the empty model is not a very good model. Very few countries have this particular value of Happiness. Even so, the mean is a pretty good prediction. Why?
3. How would we write the empty model using the GLM notation?
4. Here is a jitter plot where of **Happiness** by **spirit3group** (note: no shuffling involved), draw the empty model in this graph and also draw a few residuals.



1. Here is a jitter plot where of **Happiness** by **spirit3group** (but this time we shuffled **Happiness**), draw the empty model in this graph and also draw a few residuals.



1. Why is the empty model the same in both 30 and 31?
2. Use R to calculate the variance of **Happiness**. What is the number? What does this represent?
3. Use R to calculate the standard deviation of **Happiness**. What is the number? What does this represent? Draw the standard deviation on the histogram below.



1. What’s so great about the sum of squares?
2. What is problematic about the sum of squares? (That is, why do we need variance and standard deviation?)
3. In this jitter plot, draw a few of the “squares” that get added up in the sum of squares. Can you ever have a negative square? Why or why not?

