CSCI 36 Homework 2

Due 9/10/24

Classmates/other resources consulted: N/A

To make the figures show up smaller in the knitted file:

```
# To make the figures show up smaller in the knitted file
# This just improves readability for your graders,
# so a single figure doesn't take up almost an entire page
knitr::opts_chunk$set(fig.width=6, fig.height=3)
```

> **Throughout this assignment, you will use the following data sets. Run the following command to import and create these data sets (you will need to be connected to the internet). You do not need to know how this code chunk works.**

```
library(tidyverse)

# Load in data - must have an internet conenction for this to work
# You do not need to know how this works
london_marathon <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
london_marathon_winners <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
  mutate(Nationality2 = fct_lump_n(Nationality, 6))
species <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/202
```

Question 1 (15 points)

a. (2 points) **Output the species data set (be sure to run the code chunk before Question 1 first, to import this data set). How many rows does it have? How many columns does it have? What are those columns? To find out what the abbreviations mean, look at: https://github.com/rfordatascience/tidytuesday/tree/master/data/2023/2023-05-02 (except for commontype, which is just based on the commonname column).**

```
species
```

```
## # A tibble: 21 x 7
##    scientificname         commonname commontype granivore meanhfl meanwgt juvwgt
##    <chr>                  <chr>      <chr>          <dbl>   <dbl>   <dbl>  <dbl>
##  1 Baiomys taylori        Northern ~ Mouse              1    13.3    9.45   NA
##  2 Chaetodipus baileyi    Bailey's ~ Mouse              1    26.0   31.9    19.0
##  3 Chaetodipus hispidus   Hispid po~ Mouse              1    25.1   30.7    24
##  4 Chaetodipus intermedi~ Rock pock~ Mouse              1    22.0   17.5    10
##  5 Chaetodipus penicilla~ Desert po~ Mouse              1    21.5   17.6    11.7
##  6 Dipodomys merriami     Merriam's~ Rat                1    35.9   43.5    26.4
##  7 Dipodomys ordii        Ord's kan~ Rat                1    35.5   49.0    29.5
##  8 Dipodomys spectabilis  Banner-ta~ Rat                1    49.9  120.     76.8
##  9 Neotoma albigula       White-thr~ Rat                0    32.1  163.     83.8
## 10 Onychomys leucogaster  Northern ~ Mouse              0    20.3   31.1    18.4
## # i 11 more rows
```

```r
colnames(species)
```

```
## [1] "scientificname" "commonname"     "commontype"     "granivore"
## [5] "meanhfl"        "meanwgt"        "juvwgt"
```

Rows: 21 Columns: 7 “scientificname” “commonname”, “commontype”, “granivore”, “meanhfl”, “meanwgt”, “juvwgt”

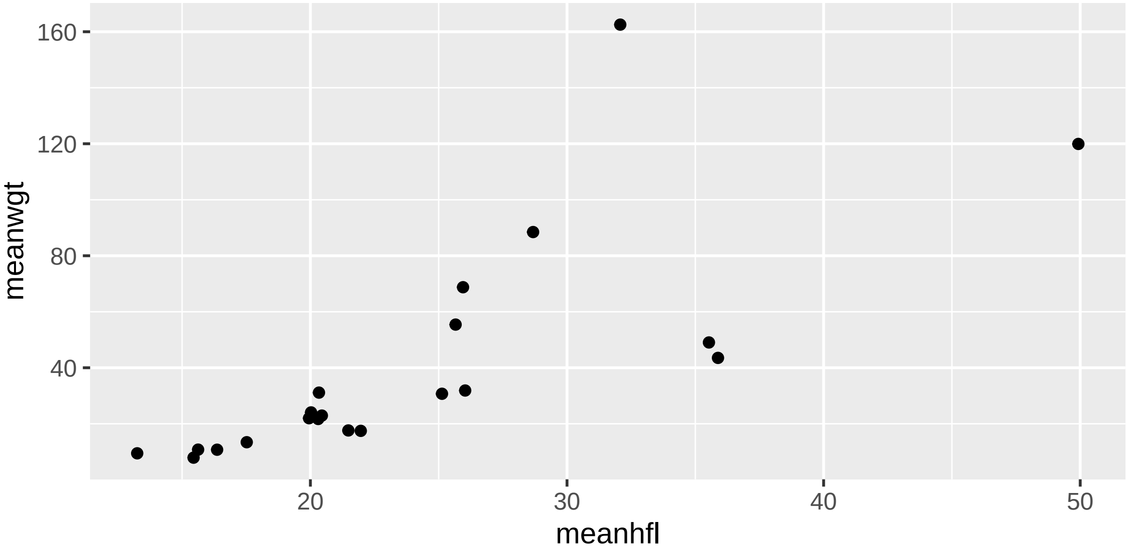
These are the descriptions from the website.

variable	class	description
species	character	Species
scientificname	character	Scientific Name
taxa	character	Taxa
commonname	character	Common Name
censustarget	double	Target species (0 or 1)
unidentified	double	Unidentified (0 or 1)
rodent	double	Rodent (0 or 1)
granivore	double	Granivore (0 or 1)
minhfl	double	Minimum hindfoot length
meanhfl	double	Mean hindfoot length
maxhfl	double	Maximum hindfoot length
minwgt	double	Minimum weight
meanwgt	double	Mean weight
maxwgt	double	Maximum weight
juvwgt	double	Juvenile weight

b. (2 points) **Make a scatterplot of mean hind foot length vs. mean weight for these species, with mean hind foot length on the x-axis and mean weight on the y-axis.**

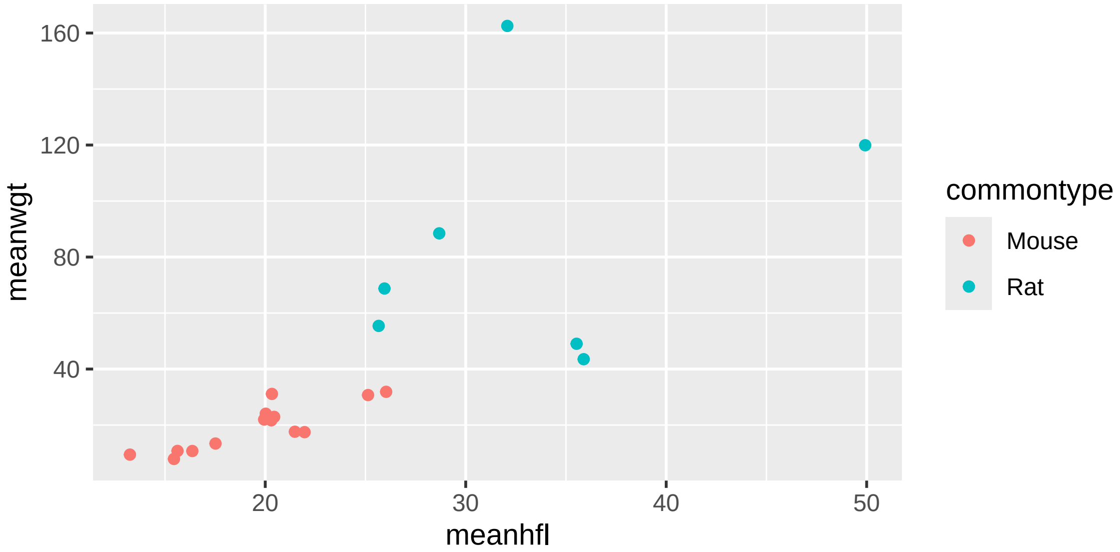
```r
library(ggplot2)

ggplot(species, aes(x = meanhfl, y = meanwgt)) +
  geom_point()
```



c. (2 points) **Add the information in the commontype variable into your scatterplot from the previous part in an appropriate way.**

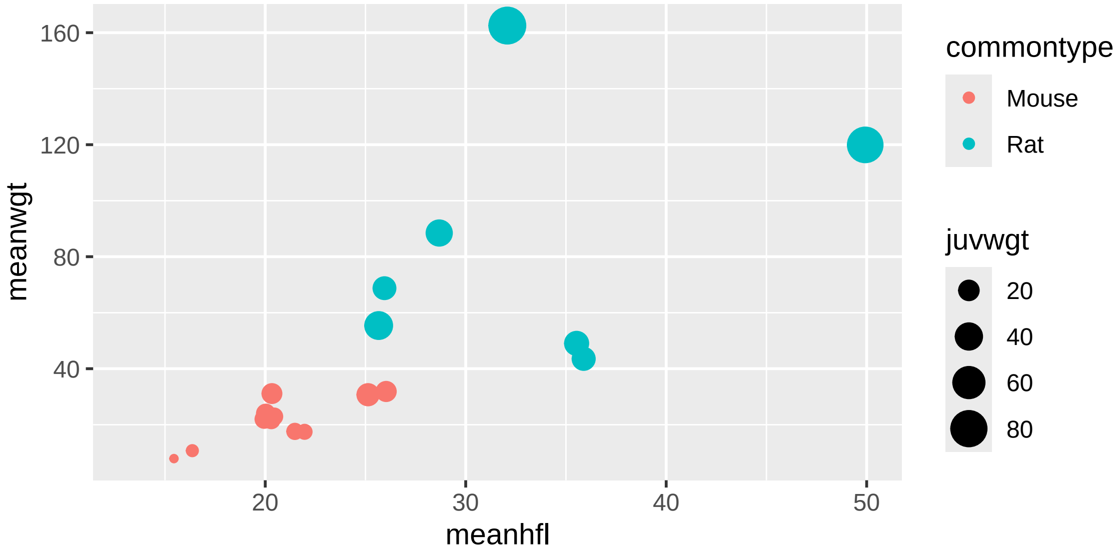
```
ggplot(species, aes(x = meanhfl, y = meanwgt, color = commontype)) +
  geom_point()
```



d. (2 points) **Add the information in the Juvenile Weight column variable into your scatterplot from the previous part in an appropriate way. CAUTION: What happens to the points corresponding to species with no juvenile weights given? Explain.**

```
ggplot(species, aes(x = meanhfl, y = meanwgt, color = commontype, size = juvwgt)) +
  geom_point()
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



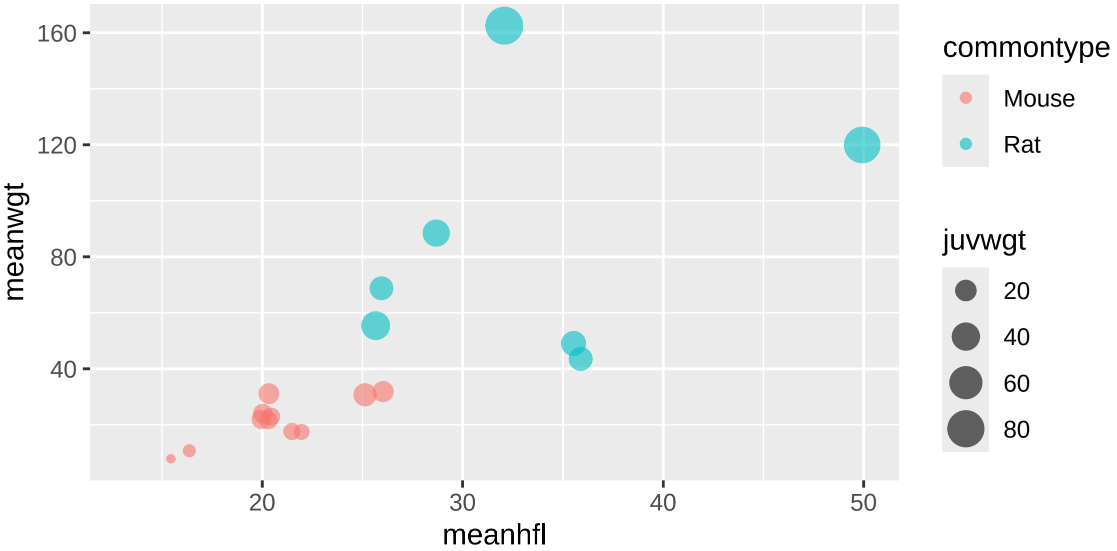
Points corresponding to species with no juvenile weights will not appear because `ggplot2` omits `NA` values by default. This can lead to missing data points and potential misinterpretation.

e. (2 points) **Modify your plot from the previous question so that your data points**

are partially transparent, so that any data points that lie on top of each other can be more easily seen.

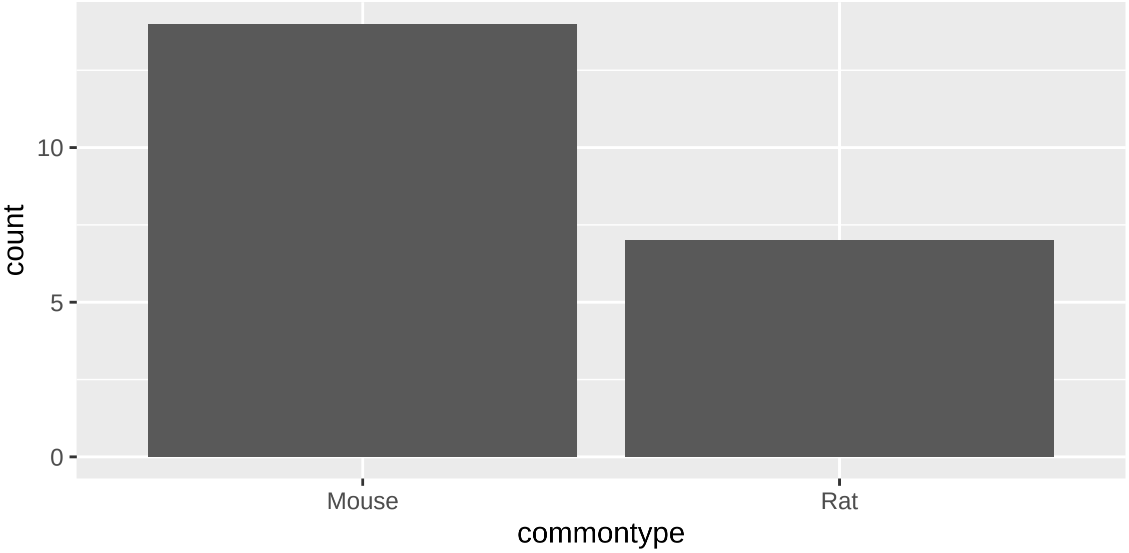
```
ggplot(species, aes(x = meanhfl, y = meanwgt, color = commontype, size = juvwgt)) +
  geom_point(alpha = 0.6)
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



f. (2 points) **Make a bar chart showing how many species in the table are commonly referred to as "mouse", and how many species in the table are commonly referred to as "rat".**

```
ggplot(species, aes(x = commontype)) +
  geom_bar()
```



g. (3 points) **Based on the plots you made in this question, make at least three observations about the relationship between hind foot length, weight, juvenile weight, and whether the species is commonly referred to as a rat or a mouse.**

1. As weight increases, hind foot length also tends to increase.

2. As weight and HFL increase, juvenile weight also tends to be larger.
3. Those with weight less than 40 and hind foot length less than 27 tend to be classified as a mouse, while those above tend to be classified as a rat.

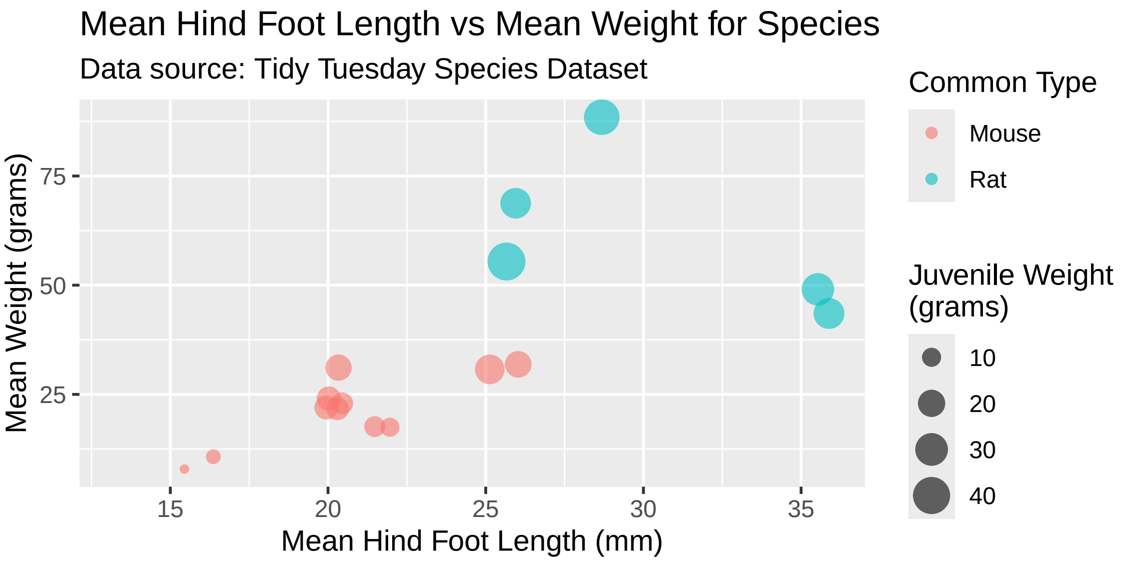
Question 2 (4 points)

For any of your scatterplots from Question 1, restrict it to only include data points where mean hind foot length is at most 40 and the mean weight is at most 100. Add a title to your plot, a subtitle describing the data source, and edit the axes labels so they are easier to understand.

```r
filtered_species <- filter(species, meanhfl <= 40 & meanwgt <= 100)

ggplot(filtered_species, aes(x = meanhfl, y = meanwgt, color = commontype, size = juvwgt)) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Mean Hind Foot Length vs Mean Weight for Species",
    subtitle = "Data source: Tidy Tuesday Species Dataset",
    x = "Mean Hind Foot Length (mm)",
    y = "Mean Weight (grams)",
    color = "Common Type",
    size = "Juvenile Weight\n(grams)"
  )
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

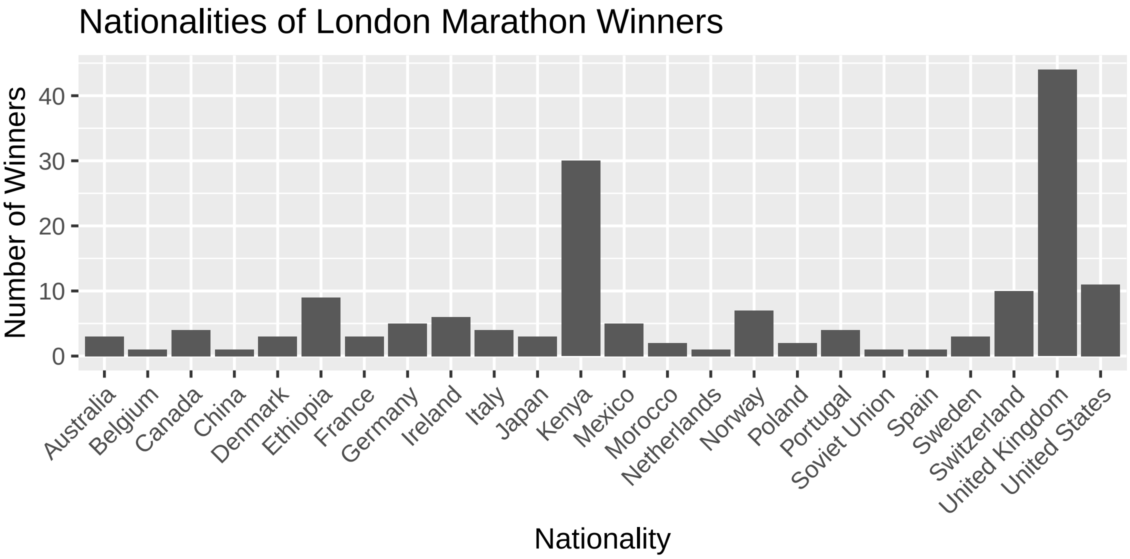


Question 3 (6 points)

a. **Using the london_marathon_winners data set, make a bar chart that shows the nationalities of the winners of the London marathon (using the Nationality column). That is, your plot should have a bar for each country represented, and the bar's size should correspond to the number of winners with that nationality.**

Make sure your chart is readable, in particular, that none of the country names overlap.

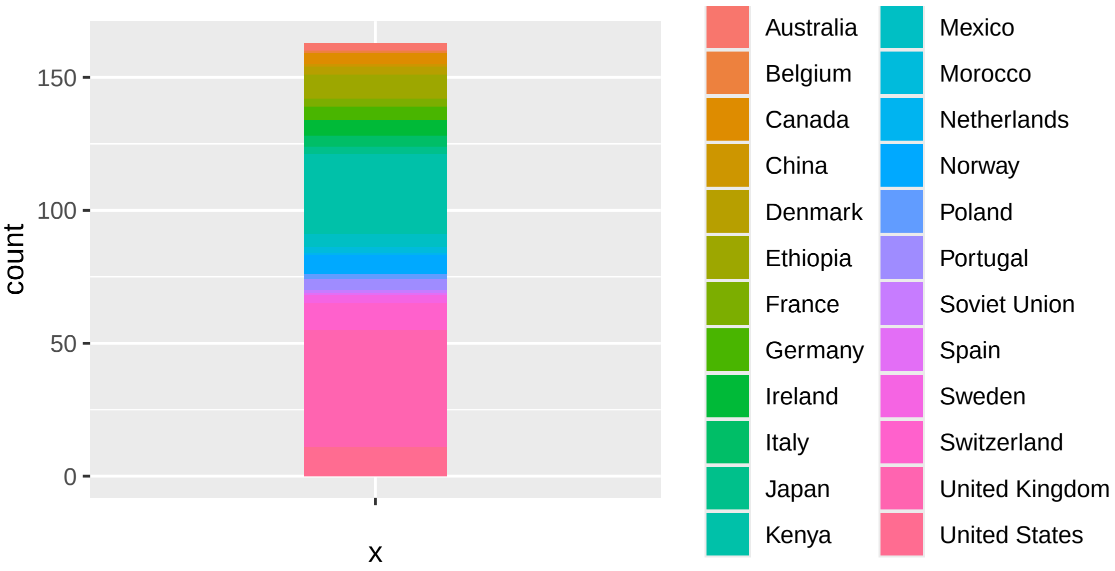
```
ggplot(london_marathon_winners, aes(x = Nationality)) +
  geom_bar() +
  labs(
    title = "Nationalities of London Marathon Winners",
    x = "Nationality",
    y = "Number of Winners"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



b. **Using the london_marathon_winners data set, make two composition plots showing the same information as the previous plot: one should use the Nationality column, and one should use the Nationality2 column. Each plot should have multiple colors but only a single bar. Which plot do you think best conveys information about the nationalities of the winners?**

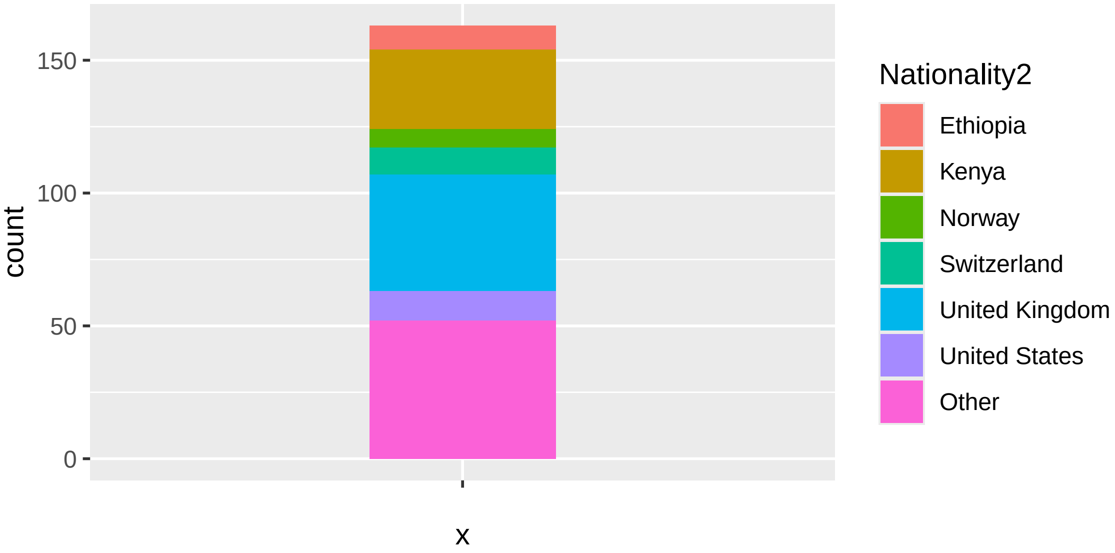
Plot 1: Using Nationality

```
ggplot(london_marathon_winners, aes(x = "", fill = Nationality)) +
  geom_bar(width = 0.3)
```



Plot 2: Using Nationality2

```r
ggplot(london_marathon_winners, aes(x = "", fill = Nationality2)) +
  geom_bar(width = 0.3)
```



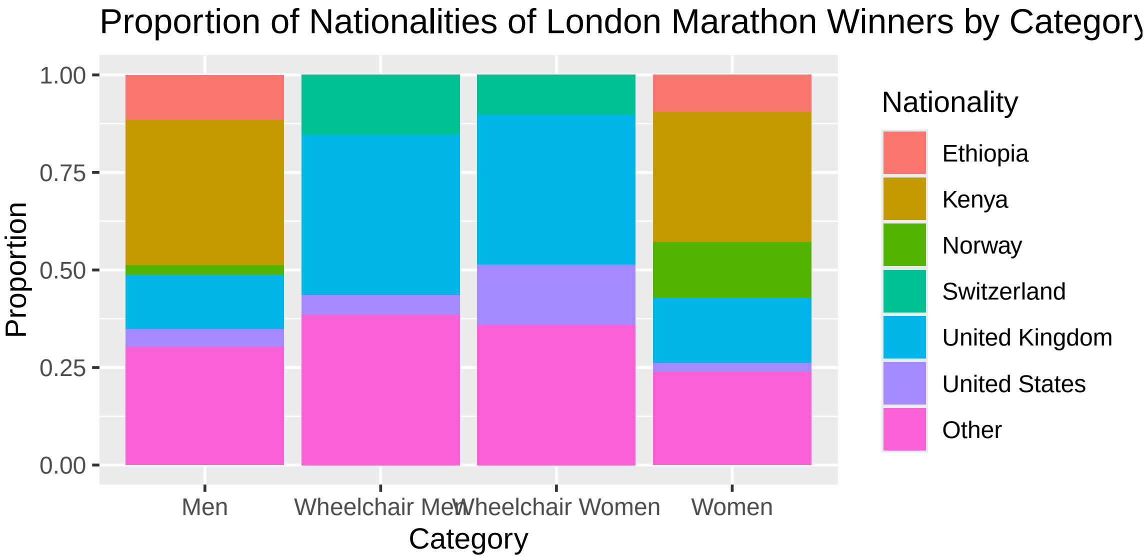
Plot 2 conveys the information more effectively by using the `Nationality2` column, which groups less frequent nationalities into an "other" category, simplifying the visualization and emphasizing the most common nationalities.

Question 4 (9 points)

In the london_marathon_winners data set, the Category column states which division the winner won: Men, Women, Wheelchair Men, or Wheelchair Women.

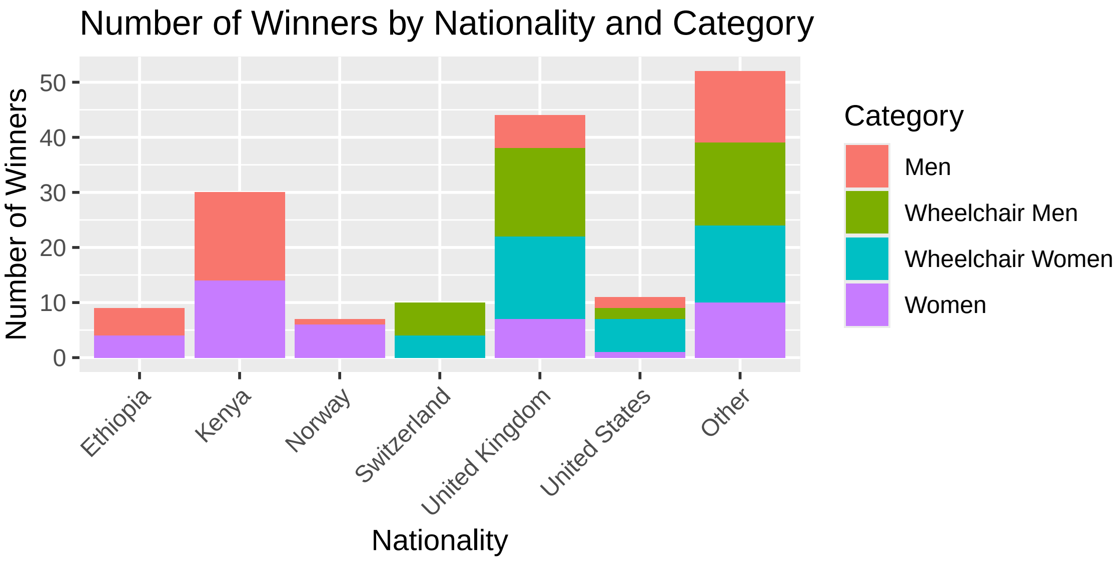
a. **Using the Category and Nationality2 columns of the london_marathon_winners data set, make a bar chart that shows, for each Category, the proportion of nationalities of the winners of that category. Be sure to relabel your y-axis appropriately. (Your chart should have four bars of the same size).**

```
ggplot(london_marathon_winners, aes(x = Category, fill = Nationality2)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion of Nationalities of London Marathon Winners by Category",
    x = "Category",
    y = "Proportion",
    fill = "Nationality"
  )
```



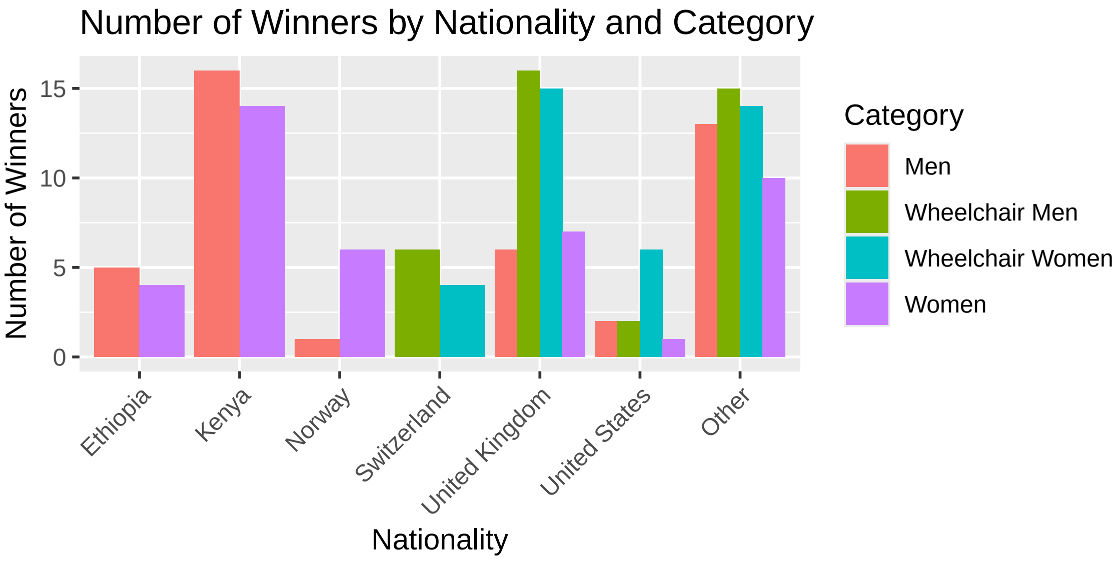
b. **Using the Category and Nationality2 columns of the london_marathon_winners data set, make a bar chart that shows, for each Nationality, how many winners of that Nationality are in each category. Make sure all text accompanying your figure is readable and non-overlapping.(Your chart should have 7 bars, of different sizes and colored differently).**

```
ggplot(london_marathon_winners, aes(x = Nationality2, fill = Category)) +
  geom_bar(position = "stack") +
  labs(
    title = "Number of Winners by Nationality and Category",
    x = "Nationality",
    y = "Number of Winners",
    fill = "Category"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



c. **Make a bar chart using the position = "dodge" option showing the relationship between the categorical variables Category and Nationality2 from the london_marathon_winners data set.**

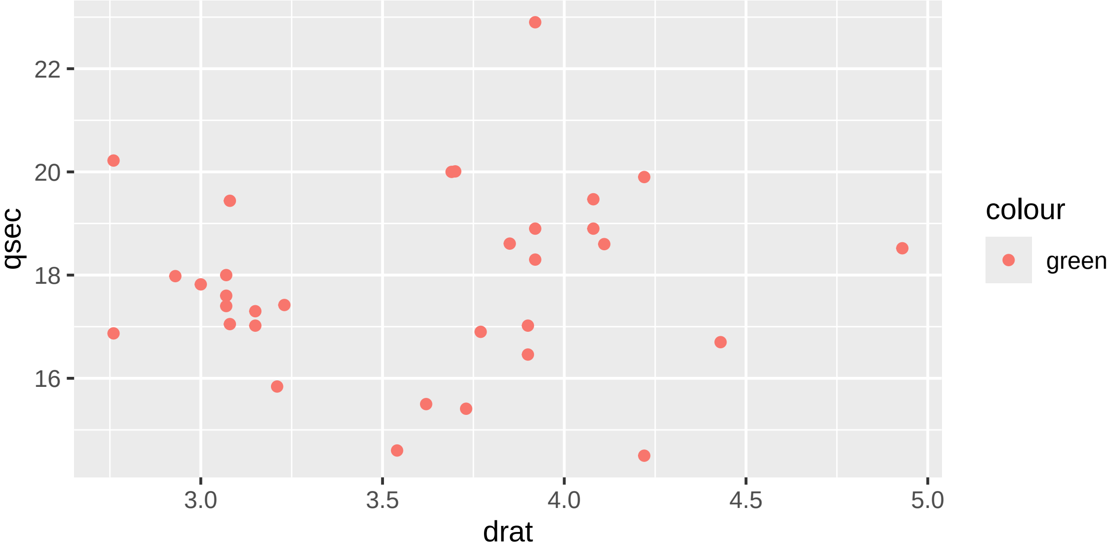
```
ggplot(london_marathon_winners, aes(x = Nationality2, fill = Category)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Number of Winners by Nationality and Category",
    x = "Nationality",
    y = "Number of Winners",
    fill = "Category"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Question 5 (9 points)

a. **What has gone wrong with this plot, and how would you change it to make all points green?**

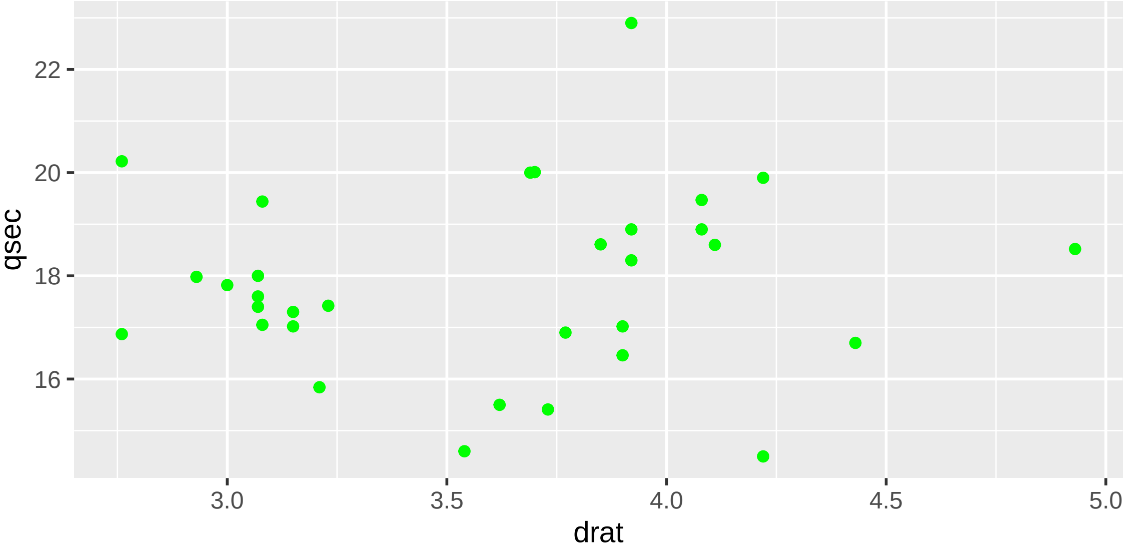
```
ggplot(data = mtcars) + geom_point(mapping = aes(x = drat, y = qsec, color = "green"))
```



In the provided plot, the color is mistakenly treated as a mapping from the data, and this means that all points are assigned to a single color category called "green" rather than being colored green.

This is the fix:

```
ggplot(data = mtcars) +
  geom_point(mapping = aes(x = drat, y = qsec), color = "green")
```



b. **Will the following two commands produce the same output? Explain why or why not.(you are welcome to run the commands and look at their output; grading will be based on your explanation).**

ggplot(data = diamonds) + geom_point(mapping = aes(x = carat, y = price))

ggplot(diamonds) + geom_point(aes(carat, price))

They will display the same. In both commands, `ggplot` is set to use the `diamonds` dataset and `geom_point` is mapping `carat` to the x-axis and `price` to the y-axis. In the first command the arguments are passed explicitly, but they don't have to be.

c. **Will the following two commands produce the same output? Explain why or why not. (you are welcome to run the commands and look at their output; grading will be based on your explanation).**

ggplot(data = diamonds, mapping = aes(x = carat, y = price)) + geom_point()

ggplot(data = diamonds) + geom_point(mapping = aes(x = carat, y = price))

In the first command, the `diamonds` dataset and the mappings (`x = carat`, `y = price`) are set in `ggplot()`. In the second command, the `diamonds` dataset is set in `ggplot()`, and the mappings (`x = carat`, `y = price`) are set in `geom_point()`. Both commands show the same scatterplot of `carat` vs. `price` using the `diamonds` dataset. The difference is only where the mappings are set, but the final plot is the same.

Question 6 (7 points)

Data Ethics: Read the article at https://viborc.com/ethics-and-ethical-data-visualization-a-complete-guide/.

a. (2 points) **According to this article, what are the most important things to keep in mind when creating a data visualization? Your answer should be about one short paragraph.**

The article says that when making a data visualization, it's important to focus on things like accuracy, simplicity, fairness, privacy, and making sure everyone can understand it. Visuals should show the data truthfully, not be misleading, and should be easy to understand. You should also think about not using personal or sensitive data without permission.

b. (2 points) **How do the ideas discussed in this article relate to what we learned in class this week? Your answer should be about one short paragraph.**

In class, we talked about how data can be misleading if not shown properly, and this article explains the same thing. It mentions how important it is to make sure visuals are honest and fair, just like we learned when discussing how people can be tricked by bad visualizations.

c. (3 points) **How do you plan to ensure in the future that any data visualizations you create follow best practices with regard to data visualization ethics? Your answer should be about one short paragraph.**

In the future, I will try to make sure my visualizations are accurate by not messing with things like scales or axes to change the way the data looks. I will also try to make my visuals simple and clear, and I'll be careful about using data that might be private. I will check to make sure my visualizations can be understood by all types of people.