36-315 Homework 1, Fall 2022

SOLUTIONS

Due Wednesday, Sept. 7, 2022 (11:59pm ET) on Gradescope

Contents

Introduction to R, RStudio, Data Types, and Critiquing Graphics

---

Problem 1: Critiquing Graphs (15 points)

As part of your course grade for 36-315, you must post on Piazza once a month describing a graph you found online (see syllabus for details). In this question, you will get practice doing this, so that expectations for these once-a-month Piazza posts are clear. Note that the following is for homework credit, NOT once-a-month Piazza credit; you CANNOT use the graphic you discuss here for your once-a-month Piazza post.

For this problem, answer the following questions:

- (5pts) First, find a graph from the Internet from the last 7 days (e.g., from a news article, blog post, online forum, etc.) or from an academic journal from the last 60 days. For this part, all you have to do is include the graph. You can either embed the graph/image directly in RMarkdown (see below for instructions), or you can include a link to the graph in your answer to this question.

Here is a link to the graph. This is from reddit.com's /r/dataisbeautiful subreddit. Following the theme of working with the Titanic data, the graph is about Leonardo DiCaprio (who starred in the Titanic movie).

- (5pts) Now **describe the graph** in 2-5 sentences. Be sure that your description touches on the following points: What does the graph show? What variables are plotted, whether it's via symbols, color, or other features of the graph? What is the main result of the graph?

The graph focuses on two variables: DiCaprio's age over time and his partner's (girlfriend's) age over time. DiCaprio's age is in yellow, and as the graph affirms, "Yup, he ages linearly," and thus DiCaprio's age is

displayed by a simple straight line; meanwhile, the ages of his partners (who change over time) are displayed in blue with bar graphs. As noted by the text in the graph, the maximum age of these partners is 25 years old. Furthermore, at the bottom of the graph, we see the picture and name of each partner over time.

From the title and subtitle of the graph, the main result is pretty clear: Although DiCaprio gets older, his partners' age tends to stay the same (early 20s).

- (5pts) Finally, **critique the graph** in 2-5 sentences. Be sure that your critique touches on the following points: What are the main goal(s) of the graph? Does the graph do a good job of achieving its goals? What are the strengths and weaknesses (if any) of the graphic? What would you change (if anything) about this graphic?

This is an excellent graph. The main goal of the graph is to demonstrate that even though DiCaprio gets older, his partners' ages tend to stay the same. The color choice of this graphic is especially excellent: All of the "DiCaprio data" is in yellow, and all of the "partner data" is in light blue, and this color choice is even consistent in the text and pictures at the bottom of the graph. In some sense, this graph just displays bivariate data (DiCaprio's age and partner's age), but there is a lot of additional "meta data" (such as the names and pictures of each partner) that adds additional nuance to this bivariate data. In this sense, even though there is a lot of "ink" in the graph, it's mostly data ink. Finally, the graph is very effective in the sense that it is very persuasive, at least for me: After seeing this graph, my respect and general like for DiCaprio has unquestionably gone down. Thus, there isn't anything in particular I would change with the graph.

Problem 2: Enroll in Piazza for 36-315 (5 points)

All questions about assignments and course material should either be asked in office hours or on Piazza. Furthermore, once a month (in September, October, and November), you will be expected to post your own "graph critique" on Piazza for the class to see (similar to Problem 1 in this homework) - see the syllabus for details. Thus, **it is critical that you enroll in the course on Piazza**, because we'll often answer questions that you or your classmates may have.

a. (0 points) If you're not already signed up, enrolled in our Piazza course here. Then, on the course Piazza page, in the top-right corner, click the Settings gear/wheel icon. Under Account & Email Settings, click Edit Email Notifications. I recommend choosing Real Time for both parts and checking the "Automatically follow every question and note" checkbox.

b. (5 points) In the syllabus for 36-315, read the section "Discussion Board: Piazza" under "Course Materials, Procedures, and Logistics". Then, write the following statement: "I certify that I have read, understood, and will follow the syllabus' rules about Piazza discussion. I also certify that I will not abuse the use of anonymous posting on the course Piazza page. Finally, I will not ask any questions that essentially ask what the answer is to an assigned problem." Then type your name. If you have any questions about rules on Piazza discussion, please email Professor Branson.

Problem 3: We were plotting data before the ship even sank (37 points)

For this problem we'll use a dataset based on the Titanic survival data, which is a famous dataset for understanding how to work with categorical variables. This dataset was obtained from Kaggle (a popular site for finding publicly-available datasets) and was modified for this assignment. Here are the data:

```r
#  Load the data into R
titanic <- read.csv("https://raw.githubusercontent.com/zjbranson/315Fall2022/master/titanic.csv")
```

a. (8pts) First, read this Kaggle description to better understand the Titanic dataset. In particular, look at the Data Dictionary and Variable Notes sections. Then, answer the following questions:

- (3pts) Use the `head()` function to display the first rows of the dataset. Based on the structure of the dataset, what does each row seem to represent (i.e., what kind of "subject" is in each row)?

```r
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           7        0      1
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                            McCarthy, Mr. Timothy J   male  54     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6            17463 51.8625   E46        S
```

Each row represents a passenger on the Titanic. For example, in each row, we can see a subject's Passenger ID and Name, suggesting that each individual passenger on the Titanic is listed per row in this dataset.

- (3pts) Name at least two categorical variables, and name at least two quantitative variables.

Survived, Pclass, Sex, Cabin, and Embarked are a few categorical variables. These are all categorical variables because each of them can be represented by a qualitative set of categories. For example, even though ***`Survived`*** *is coded as 0 or 1, this is simply representing the categories "Yes" or "No", rather than actual integers.*

Meanwhile, Age and Fare are two quantitative variables. These are both quantitative variables because they are represented with actual quantitative numbers. Age can be viewed as a discrete quantitative variable (since it is measured in whole numbers in the data) or a continuous quantitative variable (since technically someone can be 27.2 years old, although this isn't measured as such in the data). Meanwhile, Fare is certainly a continuous quantitative variable, since we can see from the data that it is measured up to several decimal places.

- (2pts) Name one ordinal categorical variable.

Pclass (passenger class) is an ordinal categorical variable. Passengers can be in first, second, or third class, which are ordered from most luxurious to least luxurious. Thus, these are qualitative categories, but they also have a natural order.

For the first question, be sure to include the line of code you used to answer this question. (**You should always include all code you used to answer a question, unless stated otherwise.**) Meanwhile, for the second and third questions, be sure to explain the reasoning for your answers. (For example, if you state that a variable is a categorical variable, be sure to explain why.)

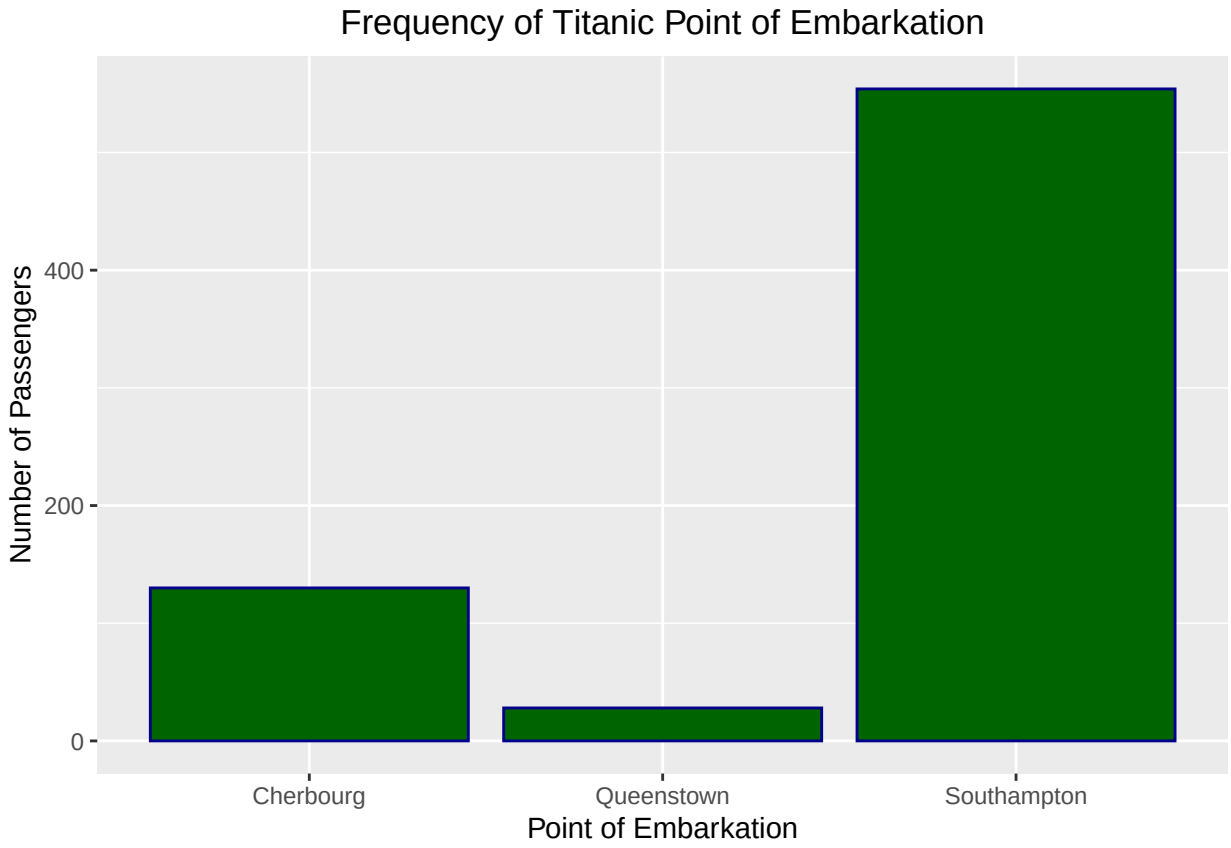
b. (10pts) For the remainder of this problem we're going to focus on the `Embarked` variable. For this part, use the `ggplot()` function to create a bar chart of the `Embarked` variable. Make sure your plot has the following characteristics:

- Its axes are properly labeled and has a proper title.
- Each bar in the graph has the same color by using `+ geom_bar(fill = "pink", color = "black")`.

After making your plot, change the `fill` and `color` arguments to something other than pink and black (feel free to choose any colors you want). Then, using your graph, describe the distribution of the `Embarked` variable in 1-2 sentences. In particular: Which port did the most passengers embark on, and which port did the fewest passengers embark on? Be sure to write your description as if it were for someone who isn't familiar with the data. **In your answer, be sure to use the actual port names, and not just "C", "Q", "S".** In general, your descriptions of graphs should make sense to people who are not familiar with the actual dataset.

First, here is the desired bar plot of the `Embarked` variable:

```
library(tidyverse)
ggplot(data = titanic, aes(x = Embarked)) +
  geom_bar(fill = "darkgreen", color = "darkblue") +
  labs(title = "Frequency of Titanic Point of Embarkation",
       x = "Point of Embarkation", y = "Number of Passengers") +
  theme(plot.title = element_text(hjust = 0.5)) +
  #make more label x-axis labels on tick marks
  scale_x_discrete(labels=c("C" = "Cherbourg", "Q" = "Queenstown", "S" = "Southampton"))
```



Here we changed the `fill` and `color` arguments from what was given in the homework (pink and black). The `fill` argument is responsible for the inner color of the bars, whereas the `color` argument is responsible for the borders of each bar.

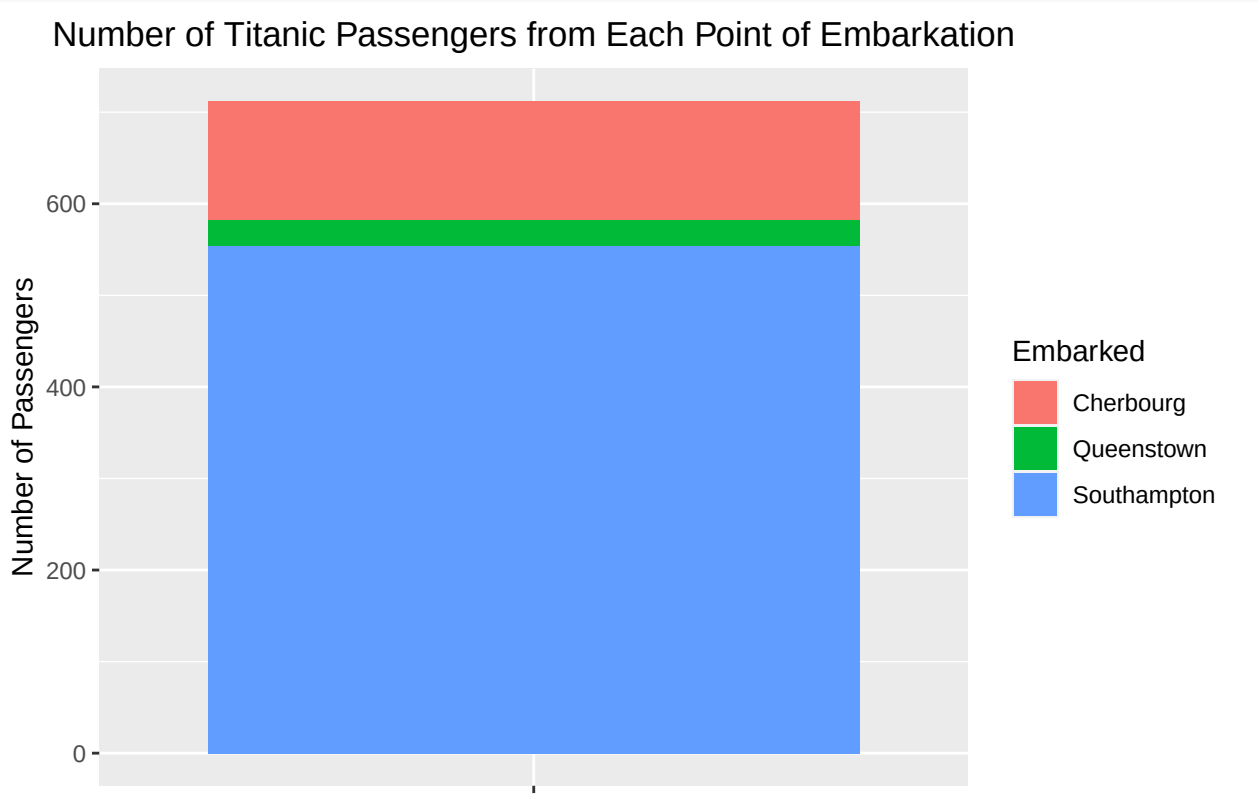
From our graph, we see that the most common point of embarkation was Southampton; over half of Titanic passengers were from Southampton. The second most common was Cherbourg; the least common was Queenstown.

[Note that we also used the `scale_x_discrete()` *function within* `ggplot` *to change the labels of the x-axis tick marks. You weren't required to do this, but it does make it easier to interpret the plot (especially for those who don't know what C, Q, and S stand for). However, you should have gotten practice using a function like this for the last question on the homework.]*

c. (8pts) For this part, answer the following questions:

- (3pts) First, make a spine chart of the `Embarked` variable. Make sure to correctly label the axes, and include an appropriate title. For this part, all you have to do is include the spine chart.

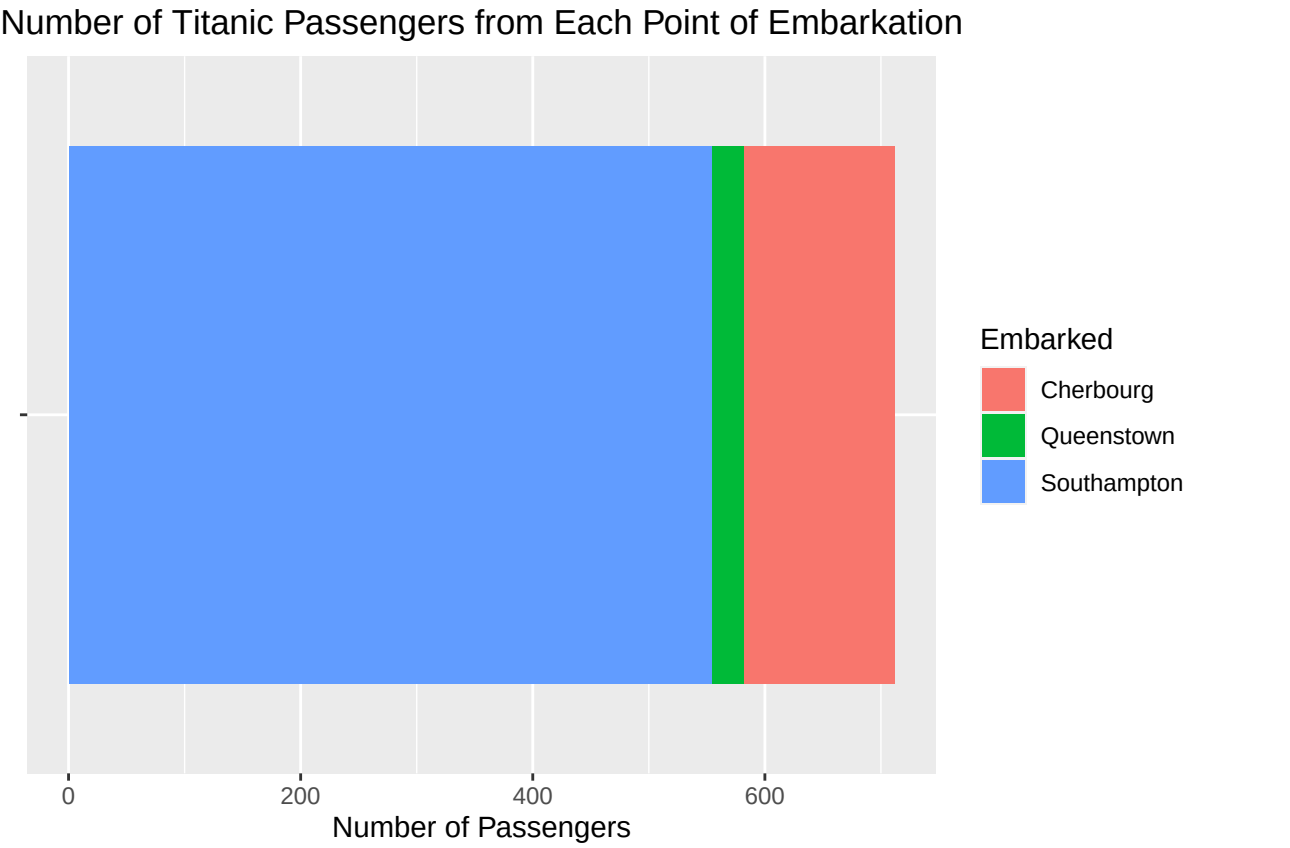
```
ggplot(data = titanic) +
  geom_bar(aes(x = "", fill = Embarked)) +
  labs(title = "Number of Titanic Passengers from Each Point of Embarkation",
       x = "",
       y = "Number of Passengers",
       fill = "Point of Embarkation") +
  theme(plot.title = element_text(hjust = 0.5)) +
  #change the legend labels
  scale_fill_discrete(name = "Embarked", labels = c("Cherbourg","Queenstown","Southampton"))
```



- (2pts) `ggplot()` allows you to easily flip the orientation of graphs without changing much of your code. To do this, you simply have to add `+ coord_flip()` to your existing code. Thus, create a new spine chart, which is exactly the same as the above spine chart, but with the orientation flipped. For this part, all you have to do is include the spine chart.

```
ggplot(data = titanic) +
  geom_bar(aes(x = "", fill = Embarked)) +
  coord_flip() +
```

```
  labs(title = "Number of Titanic Passengers from Each Point of Embarkation",
       x = "",
       y = "Number of Passengers",
       fill = "Point of Embarkation") +
  theme(plot.title = element_text(hjust = 0.5)) +
  #change the legend labels
  scale_fill_discrete(name = "Embarked", labels = c("Cherbourg","Queenstown","Southampton"))
```



- (3pts) Now answer the following: For **each** spine chart above (the first and second), what are the widths of the bars proportional to (if anything), and what are the heights of the bars proportional to (if anything)? In your answer, be sure to discuss the height and width for each of the two spine charts.

In the first spine chart, each bar has the same width (i.e., it isn't proportional to anything; or rather, it's proportional to a constant). Meanwhile, the height of each bar is proportional to the frequency of the category it represents (in this case, point of embarkation). Meanwhile, in the second spine chart, the opposite is true: The heights are constant, and the widths are proportional to the frequency of the corresponding category.

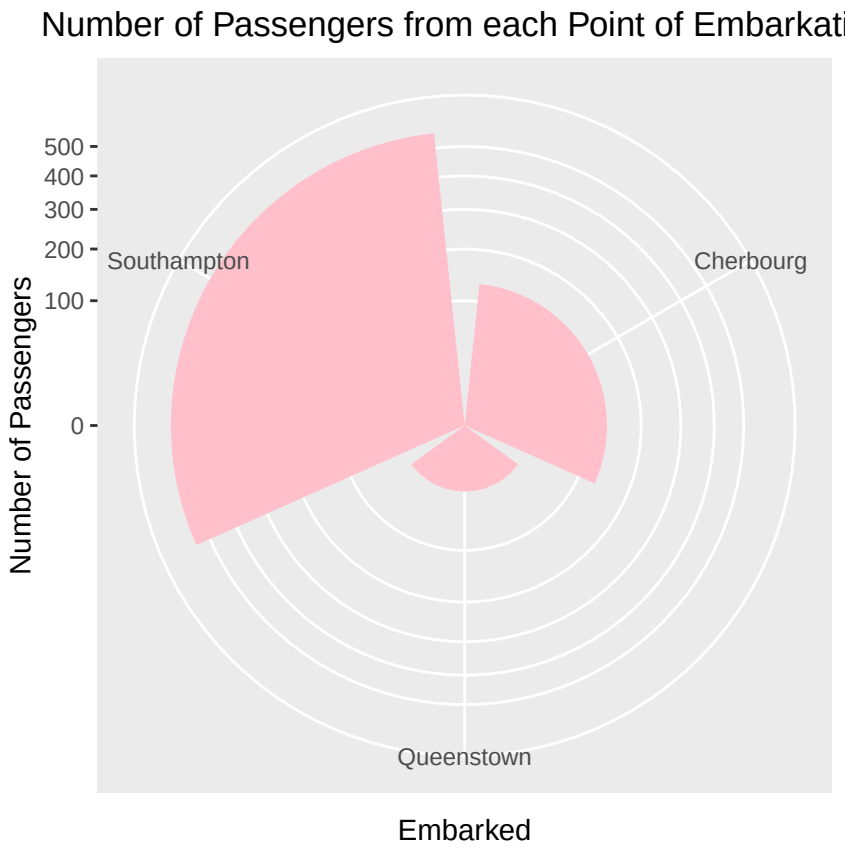
d. (6pts) Now make a rose diagram of the `Embarked` variable. Be sure to correctly label the axes and include an appropriate title. Then, answer the following questions:

- In your rose diagram, what is the radius of each rose petal proportional to?
- What does the angle associated with each rose petal correspond to (if anything)?
- What is the area of each rose petal proportional to?

Here is a rose diagram of the ***`Embarked`*** *variable:*

```
ggplot(data = titanic, aes(x = Embarked)) +
  geom_bar(fill = "pink") +
  coord_polar() + scale_y_sqrt() +
```

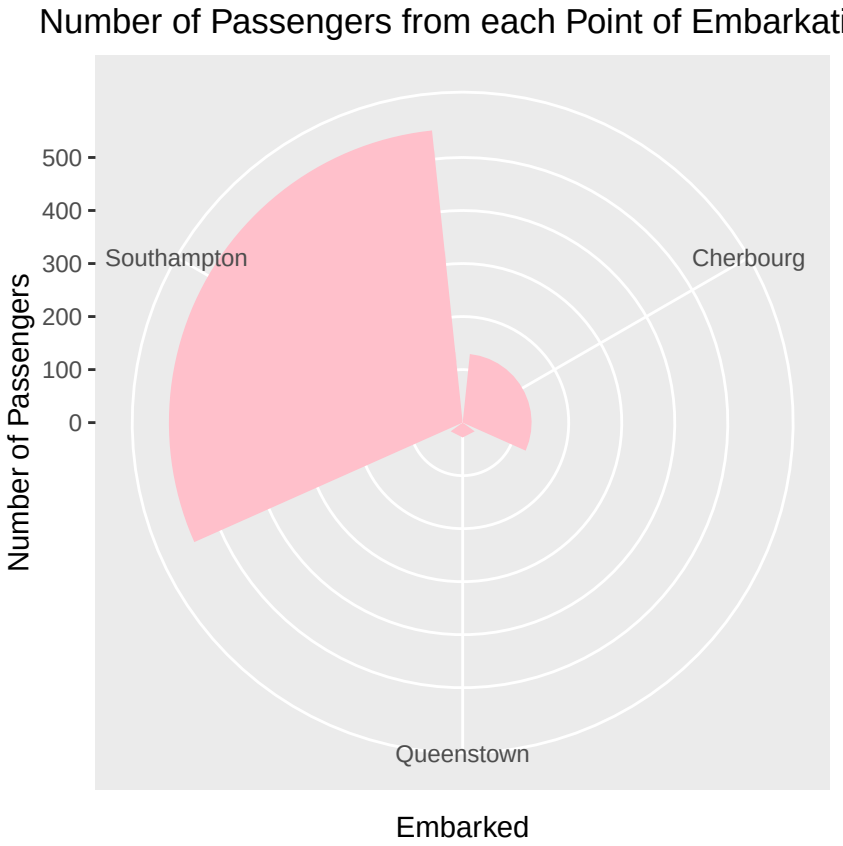
```
  labs(title  = "Number of Passengers from each Point of Embarkation",
       x      = "",
       y      = "Number of Passengers",
       fill   = "Point of Embarkation") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_discrete(name = "Embarked", labels = c("Cherbourg","Queenstown","Southampton"))
```



In our rose diagram, the radius is proportional to the square root of the number of passengers (because we used `scale_y_sqrt()`*). The angle is constant across categories (i.e., the angle is proportional to a constant). As a result, the area of each petal is proportional to the number of passengers (because the area is proportional to the square of the radius).*

Note that if you did not use `scale_y_sqrt()`*, this is what the rose diagram would look like:*

```
ggplot(data = titanic, aes(x = Embarked)) +
  geom_bar(fill = "pink") +
  coord_polar() +
  labs(title  = "Number of Passengers from each Point of Embarkation",
       x      = "",
       y      = "Number of Passengers",
       fill   = "Point of Embarkation") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_discrete(name = "Embarked", labels = c("Cherbourg","Queenstown","Southampton"))
```



Yes, Queenstown was the least frequent point of embarkation, but now the infrequency of Queenstown is quadratically exaggerated (quadratically because the area is proportional to the square of the radius, and here the radius is proportional to the number of passengers). For example, from the data we can find that Queenstown is about 20 times less frequent than Southampton; however, in the above (incorrect) rose diagram, the area for Queenstown is approximately $20^2 = 400$ *times smaller than the area for Southampton.*

e. (5pts) At this point, you've made a bar chart, a spine chart, and a rose diagram to visualize the `Embarked` variable. Which of these graphs do you most prefer, and which do you least prefer? Give a 1-3 sentence explanation as to why. In your answer, be sure to mention at least one of the "principles of data visualization" that we discussed in class. (**Hint**: There isn't necessarily one correct answer for this question—you should give your genuine, but well-reasoned, opinion.)

I most prefer the bar chart, because it allows the easiest comparison of category frequencies: I personally think it's easier to compare side-by-side heights (as in the bar chart) than stacked heights/widths (as in the spine chart) or petal areas (as in the rose diagram). In other words, I think the bar chart is the best at encouraging a comparison of different pieces of data, which is one of the data visualization principles that we discussed in class.

Meanwhile, I least prefer the rose diagram, even though it may arguably look the coolest. First, as discussed above, you can easily cause data distortion with rose diagrams, and one of the principles we discussed is to avoid data distortion. Furthermore, when looking at the inner circles in the rose diagram, I think it takes a while to figure out what they represent—i.e., it takes a while to realize that they indeed represent the numbers listed on the y-axis. In other words, when looking at the rose diagram, I have to think a bit longer about the graphical methodology, rather than the actual data—this goes against another one of the principles we discussed in class.

Problem 4: Visualizing really old data with modern ggplot2 (19 points)

In class we discussed arguably the first statistical graphic, by Langren in 1644, which visualized estimates of the longitudinal distance between Toledo and Rome. Install the `HistData` package and the `forcats` package. Then, load the `Langren1644` dataset into `R` by typing `data(Langren1644)`. See below for some example code (which also corrects a naming error in the `Langren1644` dataset). Before starting this problem, be sure to uncomment the example code below. Note that you'll have to make sure the three packages at the beginning of the code block are installed for the code to properly run.

```
#make sure these packages/libraries are installed
library(tidyverse)
library(HistData)
library(forcats)

#load the data
data(Langren1644)

#Correcting the slight misspelling of "Italy " to "Italy" in the Langren1644 data.
#The code collapses the two levels "Italy" and "Italy " of the factor variable
Langren1644 <- Langren1644 %>% mutate(Country = fct_recode(Country,"Italy" = "Italy "))
```

a. (6 points) The plots you made for the Titanic data in Problem 3 visualize a single variable; now we will visualize multiple variables within a single graph. We'll discuss the statistical nuances of graphs with multiple variables later on—the purpose of this problem is just to further explore the `ggplot()` function.

In this part, we'll create a scatterplot; I'm assuming most of you are familiar with such a plot, but we'll talk about it in depth later in the course. For now, **the goal of this problem is to make a scatterplot where**:

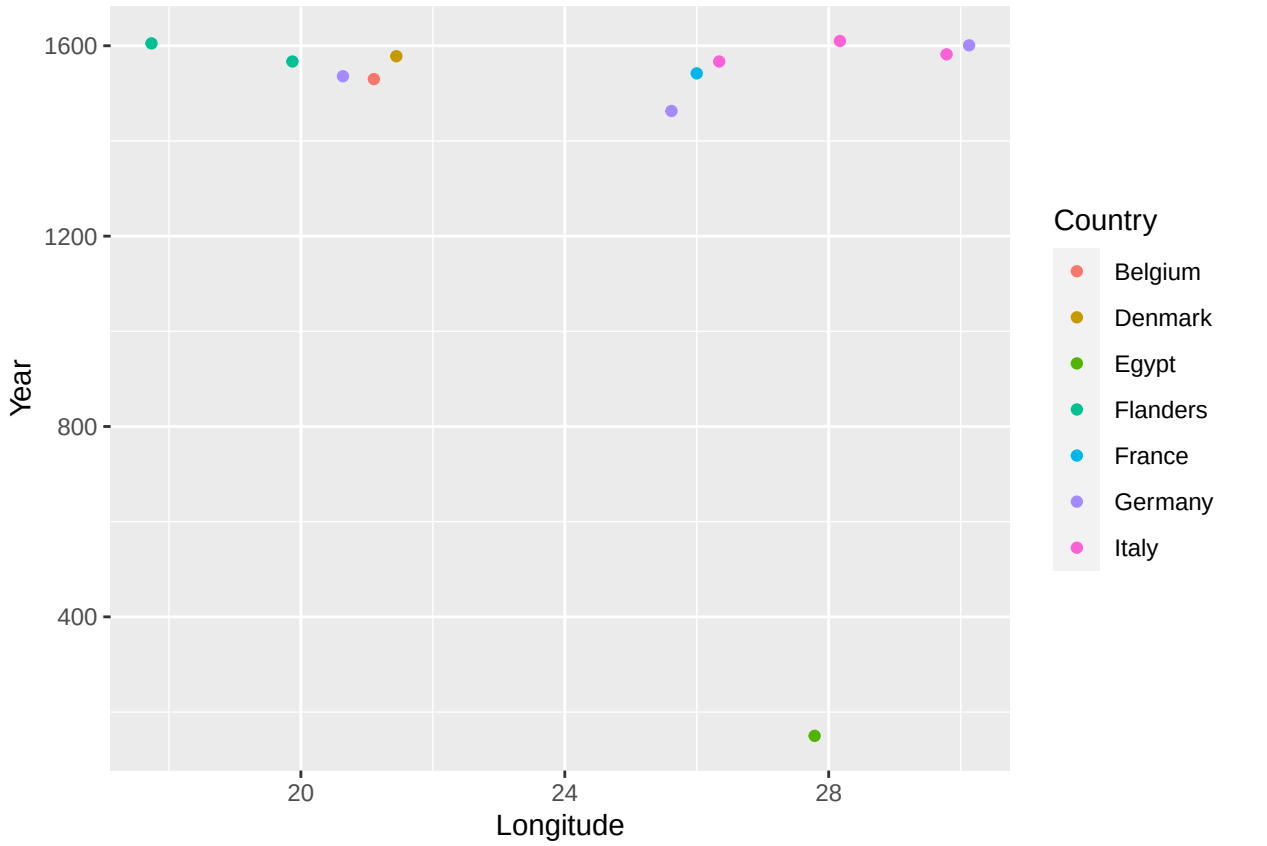
- `Longitude` is on the x-axis
- `Year` is on the y-axis
- The points in the scatterplot are colored by `Country`

To help you with this problem, I've given you some "template code" below. So, modify the below code such that you've made a scatterplot that fulfills the criteria described above.

After you've made your scatterplot, write 1-2 sentences describing what the scatterplot is showing. You may have to look at the help documentation at `help(Langren1644)` to understand what the variables in your scatterplot are.

First, here is the desired scatterplot:

```
ggplot(data = Langren1644, aes(x = Longitude, y = Year)) +
  geom_point(aes(color = Country))
```



In this graph, the x-axis refers to the estimated value of the longitude distance between Toledo and Rome, and the y-axis refers to the year that estimate was made. Furthermore, the color denotes the country associated with the estimate.

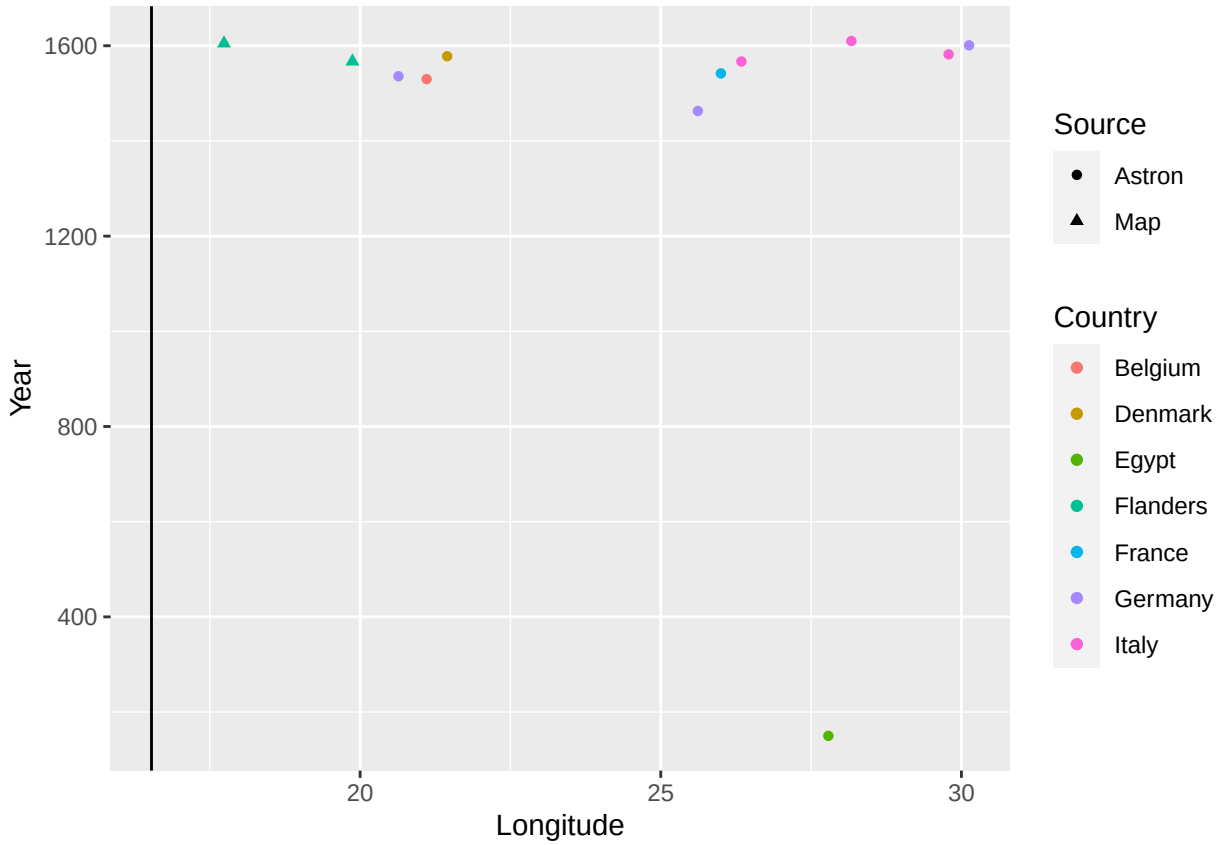
b. (5 points) Note that even though a scatterplot is considered a "two dimensional" visual, your scatterplot in Part A should have three variables (longitude, year, and country), and thus it is sort of like a "3D plot." Now we're going to make a "4D plot." **The goal of this problem is to make a scatterplot where**:

- `Longitude` is on the x-axis
- `Year` is on the y-axis
- The points in the scatterplot are colored by `Country`
- The shape of the points correspond to the **source** of longitudinal measurement.
- There's a vertical line indicating the true longitudinal distance between Toledo and Rome.

To do this problem, first copy-and-paste the code you wrote for Part A. After you do that, already the first three bullet points are done! Meanwhile, to change the shape of the points, mimic what you did with the `color` parameter in `aes`, but this time do `aes(color = ..., shape = ...)`, where `color` and `shape` are appropriately specified. Then, to add a vertical line, first use Google or `help(Langren1644)` to figure out the true longitudinal distance between Toledo and Rome. Then, use `+ geom_vline(aes(xintercept = ?))` right after your `geom_point(...)` code, where you should replace `?` with the longitudinal distance between Toledo and Rome.

For this part, all you need to do is write the code to produce the desired scatterplot.

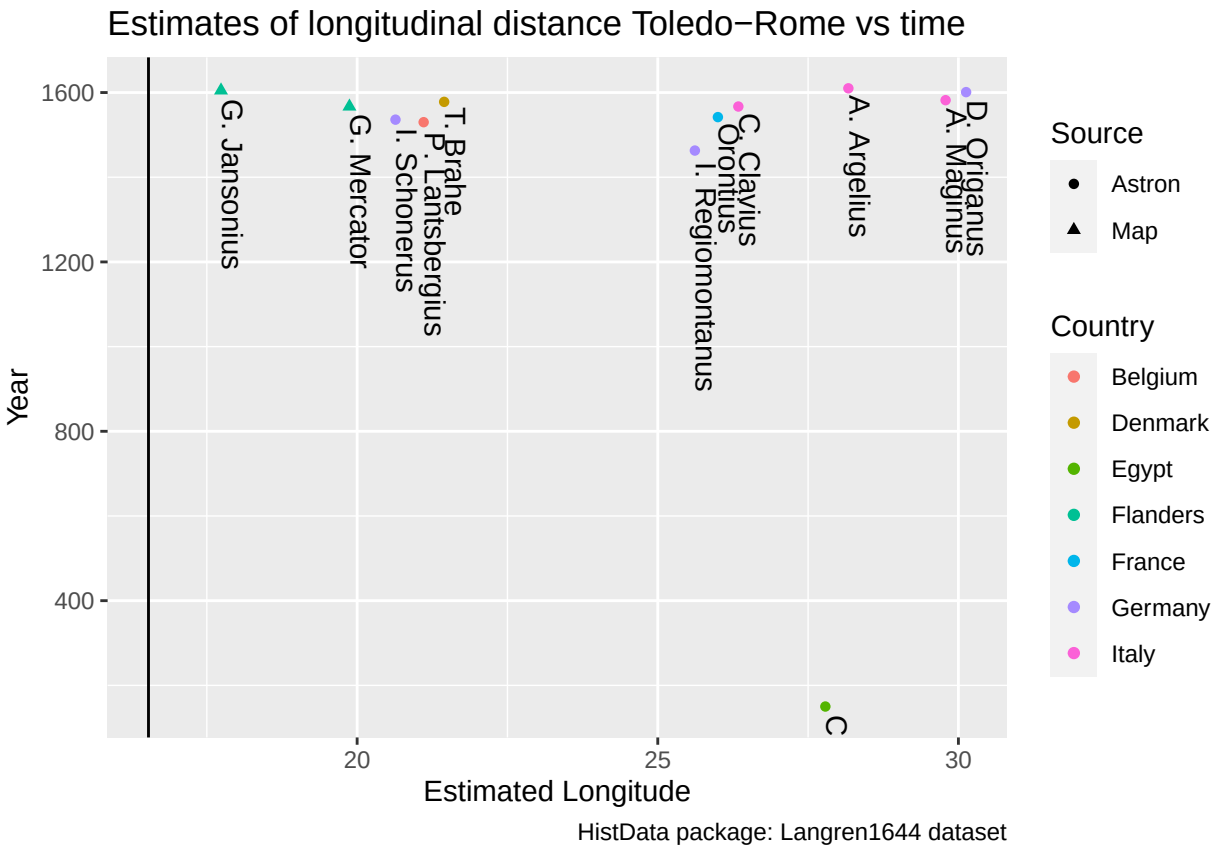
```
ggplot(data = Langren1644, aes(x = Longitude, y = Year)) +
  geom_point(aes(color = Country, shape = Source)) +
  geom_vline(aes(xintercept = 16.53))
```



c. (8 points) Now we're going to make one more scatterplot. First, copy-and-paste your code from Part B and place it here. Then, after your `geom_point(...)` code, use `+ geom_text(aes(label = Name), hjust = -0.05, vjust = -0.05, angle = -90)`. This should add text next to each point; revisit `help(Langren1644)` to understand what exactly this text is. It's okay if some of the text isn't displayed in the plot (particularly the bottom part of the plot).

After you've made your plot, add a title that (briefly) describes the plot.

```
ggplot(data = Langren1644, aes(x = Longitude, y = Year)) +
  geom_point(aes(color = Country, shape = Source)) +
  geom_vline(aes(xintercept = 16.53)) +
  geom_text(aes(label = Name), hjust = -0.05, vjust = -0.05, angle = -90) +
  labs(
    title = "Estimates of longitudinal distance Toledo-Rome vs time",
    x = "Estimated Longitude",
    caption = "HistData package: Langren1644 dataset"
  )
```



Then, answer the following questions:

- Who gave the most accurate estimate?

G. Jansonius gave the most accurate estimate.

- Which country gave the most accurate estimates?

The most accurate estimates were made by individuals from Flanders.

- Is the oldest estimate the worst?

The oldest estimate is not the worst, as there were 3 later estimates, made in Italy and Germany, that are worse than the oldest estimate.

- Which source seems more accurate?

The map seems to be more accurate, since the two most accurate estimates used maps.

Problem 5: Variable Types (24 points)

In this question, we will consider a toy dataset of 100 people with the following variables:

- `hair`: The hair color of each person (black, brown, blonde, red, or other).
- `age`: Number of years this person has been alive
- `income`: Measured in US dollars.
- `children`: Number of children this person has.

- `opinion`: Response to the statement: "There shouldn't be any 36-315 homework." Here, 3 = agree, 2 = not sure, 1 = disagree.

The toy dataset can be viewed here. Here's the code to load in the dataset:

```
data <- read.csv("https://raw.githubusercontent.com/zjbranson/315Fall2022/master/hw1DataTypes.csv")
```

a. (9pts) For this part, answer the following questions:

- (3pts) For each of the five variables above, state whether the variable is categorical (nominal), categorical (ordinal), quantitative (discrete), or quantitative (continuous).

Below are the variable types for each variable in the toy dataset:

- *`hair`: Categorical (nominal)*
- *`age`: Quantitative (continuous). Here we wrote "continuous" because it's possible for someone to be, e.g., 40.5 years old. However, there are only whole numbers in the dataset, so it's fine if you wrote "discrete" instead.*
- *`income`: Quantitative (continuous)*
- *`children`: Quantitative (discrete)*
- *`opinion`: Categorical (ordinal)*
- (2pts) For each of the five variables above, state the class of the variable according to `R` using the `class()` function. (**Hint**: A demonstration of how to do this for the `hair` variable is below. We can see that the class of the `hair` variable is "character", so already you have one of the answers! Also: For some versions of `R`, the class for `hair` may appear as `"factor"` instead of `"character"`, which is fine. Write similar code for the other variables.)

```
# #classes of the variables
class(data$hair)
```

```
## [1] "character"
```

```
class(data$age)
```

```
## [1] "integer"
```

```
class(data$income)
```

```
## [1] "numeric"
```

```
class(data$children)
```

```
## [1] "integer"
```

```
class(data$opinion)
```

```
## [1] "integer"
```

Below are the `R` classes for each variable in the toy dataset:

- *`hair`: "character"*
- *`age`: "integer"*
- *`income`: "numeric"*
- *`children`: "integer"*
- *`opinion`: "integer"*
- (4pts) Given your understanding of `R` classes and quantitative/categorical variables (as we discussed in class), which variable(s) in the dataset have unintuitive classes? Explain your answer in 1-2 sentences.

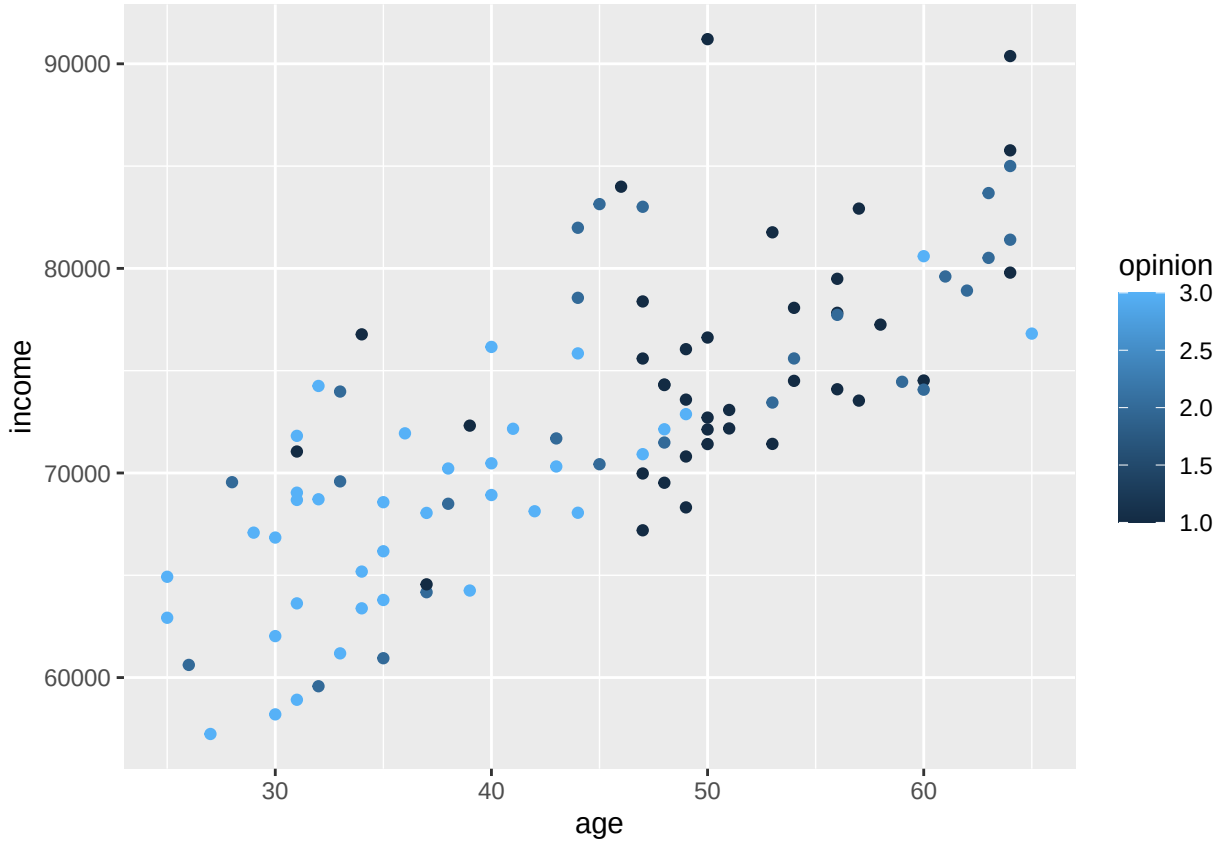
The variable ***opinion*** *has an unintuitive class; it has class "integer" (suggesting that it is a quantitative variable), but in fact it is a categorical variable. It is true that* ***opinion*** *is an ordinal variable, but nonetheless the "integer" class gives the false impression that* ***opinion*** *is quantitative when it is not.*

b. (10pts) For this part, answer the following questions:

- (4pts) First, use `ggplot()` to make a scatterplot where `age` is on the x-axis, `income` is on the y-axis, and points are colored by `opinion`. You should be able to use code very similar to Question 4 to do this. Your scatterplot should have a color legend for the `opinion` variable on the right-hand side. From this legend, how many possible values does there appear to be for the `opinion` variable?

Here is the desired scatterplot:

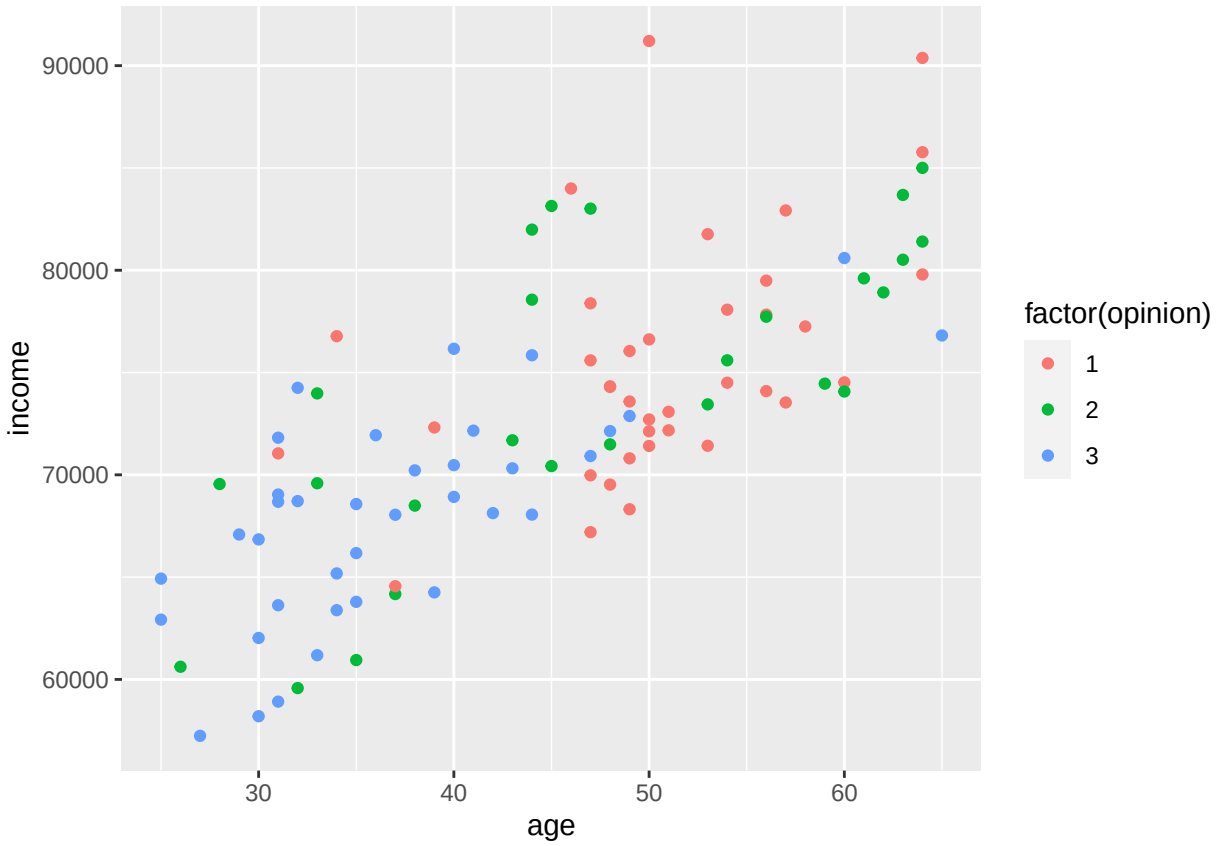
```
ggplot(data = data, aes(x = age, y = income)) +
  geom_point(aes(color = opinion))
```



From this legend, it seems like there are at least five possible values for the ***opinion*** *variable: 1.0, 1.5, 2.0, 2.5, and 3.0. In fact, it suggests that there are an infinite number of values (any number between 1 and 3).*

- (3pts) It's often helpful to change the class of variables using functions like `as.numeric()` and `as.factor()`. Luckily, `ggplot()` makes it very easy to convert variables to factors: Simply place `factor()` "around" a variable name within `ggplot()`. So, for this part, make the same scatterplot from the previous bullet point, but with `opinion` converted to a factor by using `factor(opinion)`. Your scatterplot should still have a color legend for the `opinion` variable on the right-hand side. From this legend, how many possible values does there appear to be for the `opinion` variable?

```
ggplot(data = data, aes(x = age, y = income)) +
  geom_point(aes(color = factor(opinion)))
```



From this legend, it seems like there are only three possible values for the ***opinion*** *variable: 1, 2, or 3.*

- (3pts) Now you've made two scatterplots; which scatterplot do you think is more intuitive? Explain your answer in 1-2 sentences.

The second scatterplot is more intuitive, because indeed 1, 2, or 3 are the only possible values for the ***opinion*** *variable.*

c. (5pts) Let's focus a bit more on the second scatterplot in Part B, where you used `factor(opinion)`. The color legend probably looks a bit ugly, in the sense that people not familiar with this dataset may have trouble understanding it. For this part, first copy-and-paste your code from the second bullet point in Part B and place it here. Then, to make the color legend more understandable, make the following changes to your plot:

- Notice that your color legend displays "factor(opinion)", which doesn't look too great. Change this to "Opinion" instead of "factor(opinion)".
- Notice that your color legend just has the labels 1, 2, and 3, which readers not familiar with this dataset won't understand. Change these labels to something more meaningful based on the definition of `opinion` given at the beginning of this problem.

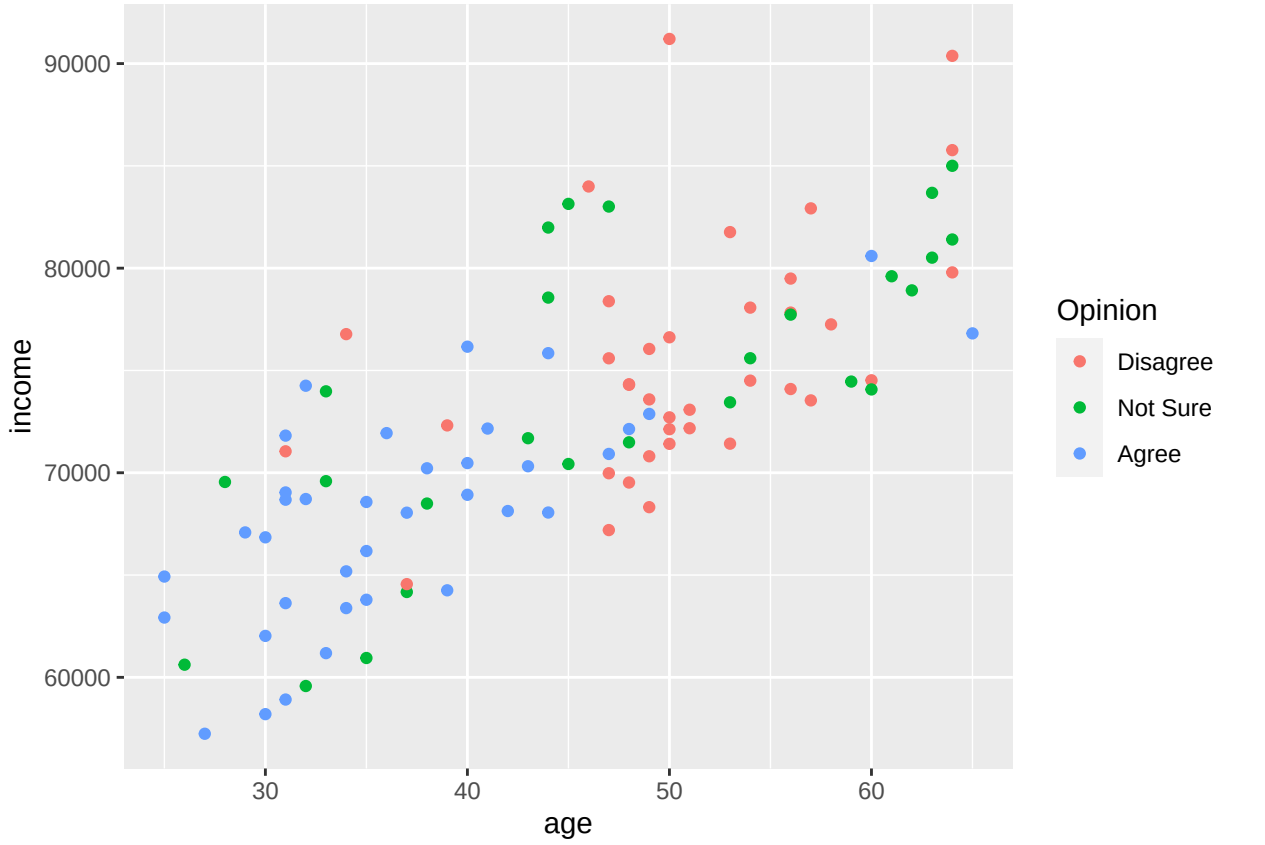
For this part, you just need to make the desired scatterplot.

Hint: For this problem I thought about just expecting you to Google how to do this and figure it out yourself, but after some reflection I thought that was too mean, especially because it's the beginning of the semester. After I Googled "change legend labels in r", the first result was this page, which I use a lot for `ggplot()` legends. See the "Rename legend labels and change the order of items" section of that page. Since you specified `color` in your plot, you'll need to use the `scale_color_discrete()` function.

Below is the desired scatterplot. Note that we used the ***scale_color_discrete()*** *function to change the legend labels for this plot. In general, there are many functions in the form of* ***scale_x_discrete()****, and*

here we use `scale_color_discrete()` *because we want to change the labels for the* `color` *argument that we specified in* `aes()`.

```
ggplot(data = data, aes(x = age, y = income)) +
  geom_point(aes(color = factor(opinion))) +
  scale_color_discrete(name = "Opinion",
    labels = c("Disagree", "Not Sure", "Agree"))
```



More generally, the point of this whole problem was to demonstrate that R can give unintuitive graphs and analyses if variables aren't appropriately classified. For this reason, it's good practice to check the class of each variable before you start making graphs and analyses. Variables like `opinion` *(where numbers are used to denote categories) are very common, so you should especially be on the lookout for those! As we will discuss later in the course, standard statistical analyses will give you inappropriate/incorrect results if you just plugged in* `opinion` *as is before recoding it as a categorical variable using a function like* `factor()`.