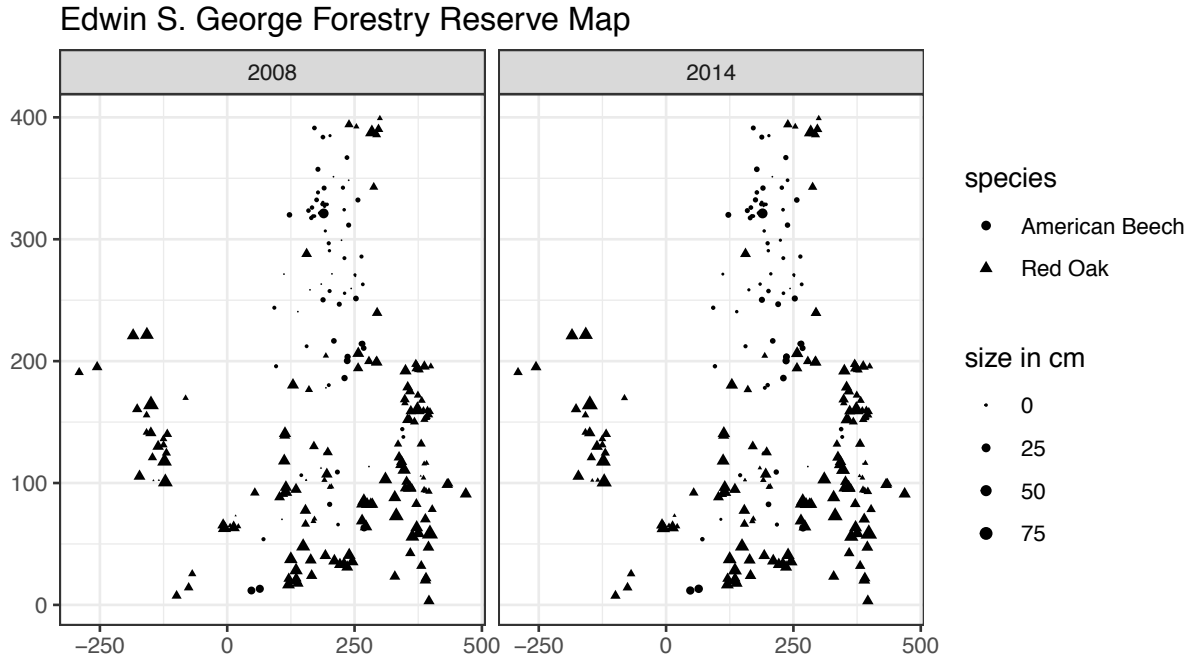
1 Graphics

Say you are walking around the Forestry Department at the University of Michigan and you see the following visualization posted on a bulletin board.

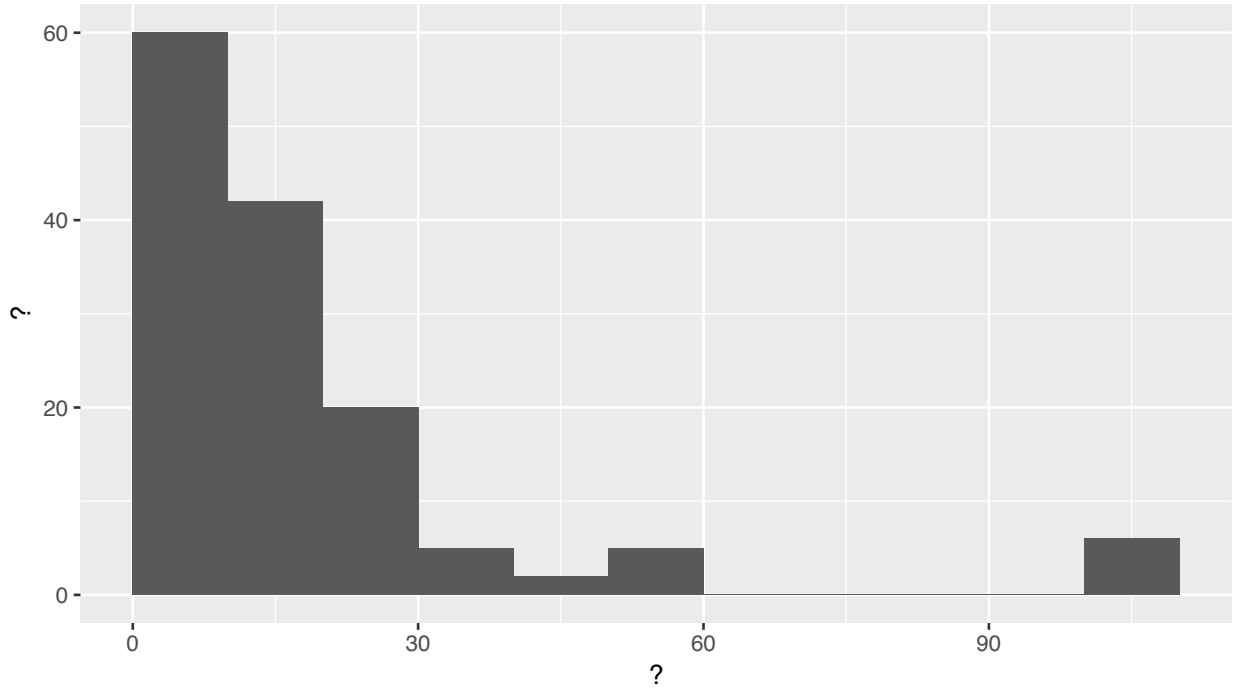


a) The geometric object of this visualization is "points." Map all the aesthetic attributes of this visualization to variables of a hypothetical dataset as is required by the "grammar of graphics" and identify all other additional components that you need to specify to create this visualization. You may treat the title of the visualization and the legends as given.

b) What is the chief comparison being made in this plot? Answer in **one sentence**.

2 Histograms

In a statistics class with 140 students, the professor records how much money (in dollars) each student has in their possession during the first class of the semester. The histogram shown below represents the data they collected:



a) What is the variable on the horizontal (x) axis?

b) What is the quantity on the vertical (y) axis?

c) The height of the second bar is 42. What does that tell us? Say precisely in **one sentence**.

d) Fill in the blanks: The median amount of money possessed is between \$ _____ and \$ _____. Show work or briefly explain your reasoning.

3 House Prices

The asking prices (in thousands of dollars) for a sample of 15 cars currently being sold are listed below. For convenience, the data have been ordered:

0	11	20	25	25	30	30	30	30	35	35	35	35	39	39

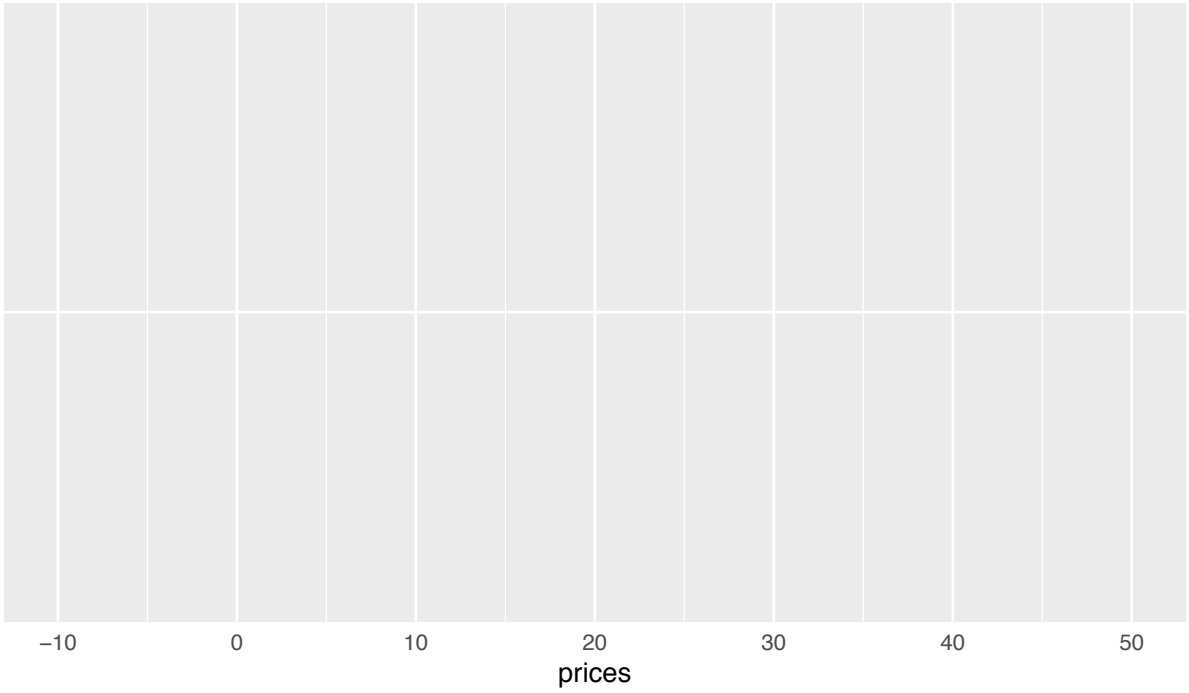
Furthermore, the following three *summary statistics* have been computed:

1st quartile	2nd quartile	3rd quartile
25	30	35

a) What is the interquartile range (IQR) for this data?

b) Is the IQR a measure of center or a measure of spread of a numerical variable? Circle your response.

c) Draw the boxplot for this dataset:



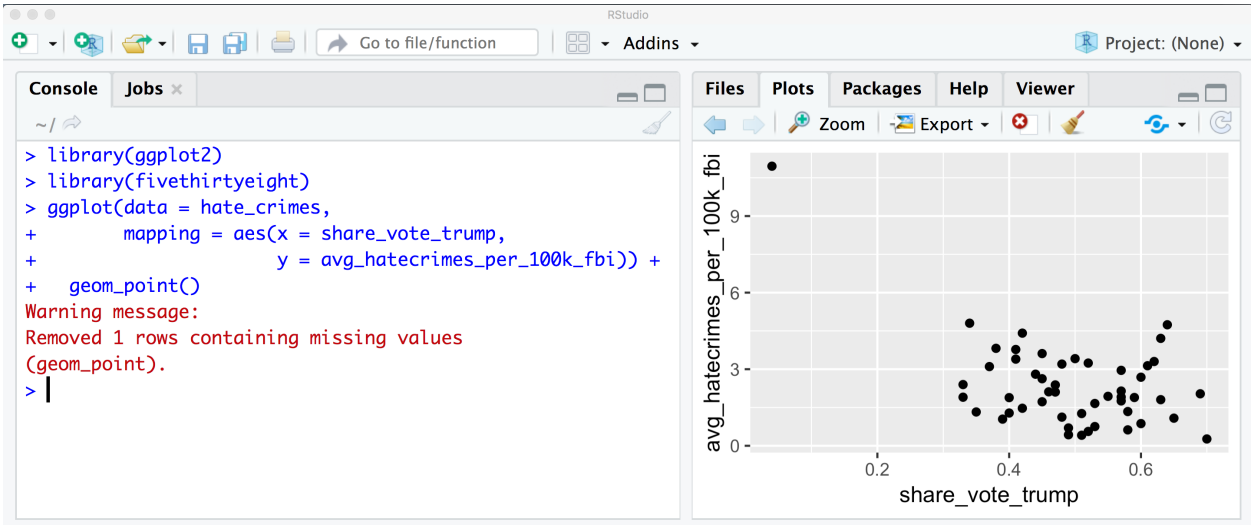
4 Short Answer

a) Say a class of 30 introductory statistics students took a test where the median score was 17 and the interquartile range was 0. What are the 25th and 75th percentiles of test scores? Hint: A picture may help.

b) Recall in Problem Set 02 you created the following scatterplot visualizing the relationship between

- the proportion that voted Trump in the 2016 election
- the average annual hate crimes per 100,000 population between 2010-2015 as reported by the FBI

for the 50 states and the District of Columbia (DC). Here is what RStudio looks like after running the necessary code in the Console:



Other than by counting the number of points, based on the above output how can we know the number of points that are in above plot?

c) What is this number of points in the above plot?

d) For which of the following pairs of variables would you visualize with a scatterplot? Circle which pairs.

- Pair 1: "Distance from school in miles" and "mode of transportation to school (bike, walking, bus)"
- Pair 2: "Number of years at a job" and "Salary"
- Pair 3: "Years experience playing an instrument" and "number of mistakes made playing a song"
- Pair 4: "Number of years since a person retired" and "favorite sport"

5 Data Sets

a) The `weather` data set in the `nycflights13` package contains hourly meterological data for the three NYC airports (EWR, JFK, and LGA); the first 6 rows of the data set are displayed below. What variables are needed to uniquely identify each observation?

origin	year	month	day	hour	temp	humid	wind_speed	precip	pressure	visib
EWR	2013	1	1	1	39	59	10.4	0	1012	10
EWR	2013	1	1	2	39	62	8.1	0	1012	10
EWR	2013	1	1	3	39	64	11.5	0	1012	10
EWR	2013	1	1	4	40	62	12.7	0	1012	10
EWR	2013	1	1	5	39	64	12.7	0	1012	10
EWR	2013	1	1	6	38	67	11.5	0	1012	10

b) Write down the arithmetic operation you would enter into a calculator to compute the number rows that the `weather` data set has. An example of an arithmetic operation is $10 \times 7 + 6$.

1 Exploratory data analysis via data wrangling

Recall the Google Forms survey you completed in Lecture 2 where:

- Students with an odd birthday (Ex: Nov 15th) were first asked if there are more or less than **14 countries** in Africa and then asked to guess how many countries there are in Africa.
- Students with an even birthday (Ex: Nov 14th) were first asked if there are more or less than **94 countries** in Africa and then asked to guess how many countries there are in Africa.

Let's refer to the numbers 14 and 94 as "priming" numbers since survey participants were "primed" with them in order to influence the number of countries they guessed. Furthermore all students were also asked their height (in inches), their graduation year (2019, 2020, 2021, or 2022), and whether or not they had previously been to Africa. A total of 41 students responded and the results are saved in a data frame `africa` with 41 rows:

```
## # A tibble: 41 x 5
##     year height been_to_africa priming      how_many_countries
##    <int>  <int> <chr>          <chr>                     <int>
##  1  2021     70 No             14 countries                 36
##  2  2020     67 No             94 countries                120
##  3  2021     69 No             14 countries                 30
##  4  2021     60 Yes            14 countries                 64
##  5  2021     66 No             14 countries                  1
##  6  2021     66 No             14 countries                 22
##  7  2022     65 No             14 countries                 16
##  8  2021     64 No             94 countries                100
##  9  2022     68 No             94 countries                 29
## 10  2021     62 No             94 countries                110
## # ... with 31 more rows
```

a) Write the pseudocode that will allow you to wrangle `africa` to obtain the median number of countries guessed for each of the two priming groups:

```
## # A tibble: 2 x 2
##   priming      median_guess
##   <chr>               <dbl>
## 1 14 countries           28
## 2 94 countries           57
```

b) Write the pseudocode that will allow you to wrangle `africa` to obtain only the year, priming group, and number of countries guessed for only the first-year students (class of 2022):

```
## # A tibble: 4 x 3
##    year priming      how_many_countries
##   <int> <chr>                     <int>
## 1  2022 14 countries                 16
## 2  2022 94 countries                 29
## 3  2022 14 countries                 30
## 4  2022 14 countries                 27
```

c) Write the pseudocode that will allow you to wrangle `africa` so that the rows are reordered from the largest number of countries guessed to the smallest (note we only show the first 5 out of 41 rows in the output below):

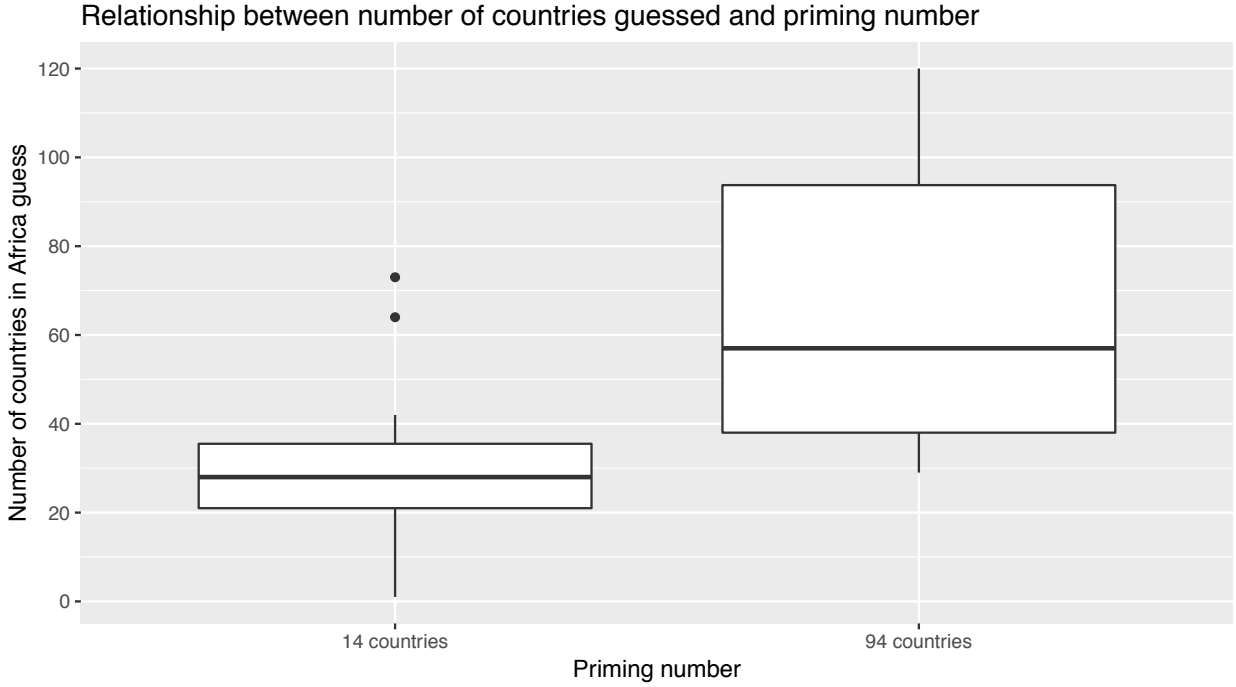
```
## # A tibble: 5 x 5
##    year height been_to_africa priming      how_many_countries
##   <int>  <int> <chr>          <chr>                     <int>
## 1  2020     67 No             94 countries                120
## 2  2021     62 No             94 countries                110
## 3  2021     64 No             94 countries                100
## 4  2021     60 No             94 countries                100
## 5  2019     68 No             94 countries                 75
```

2 Exploratory data analysis via visualizations

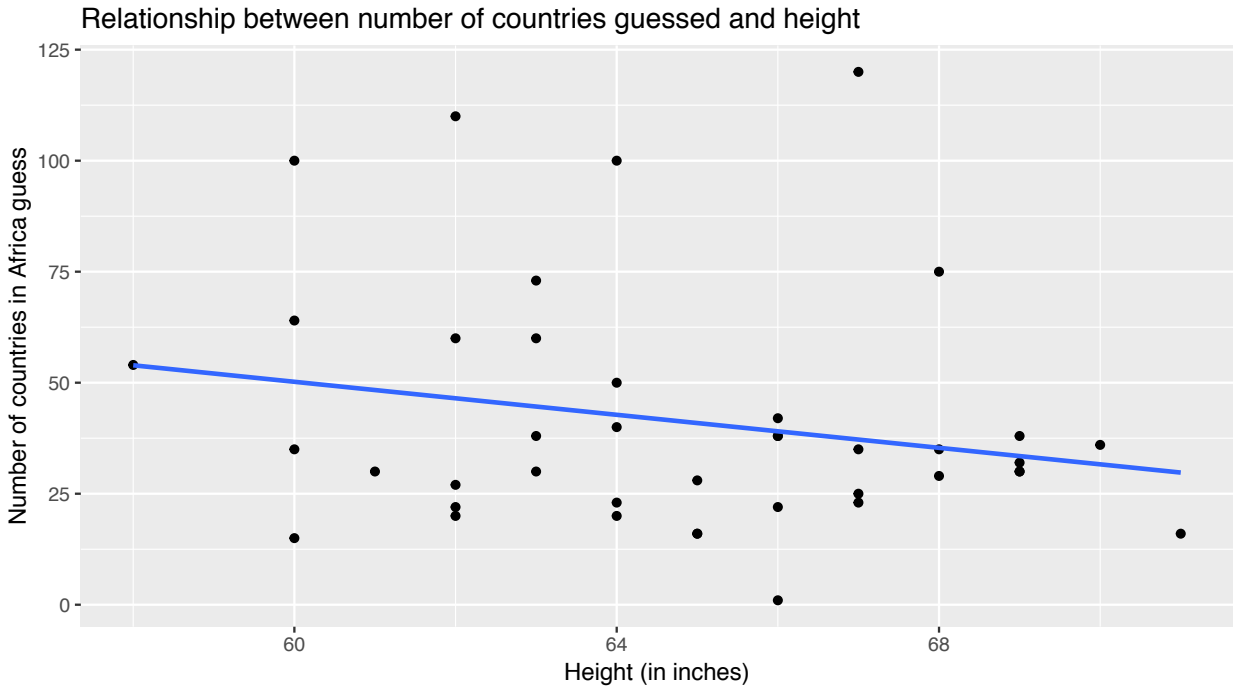
Continuing the previous `africa` question, for the remainder of this midterm let the outcome variable y be the number of countries a student guesses.

a) Name an ideal exploratory data visualization for the relationship between y and `height`.

b) We present an exploratory boxplot of the relationship between y and `priming`. It is a fact that there is more variation in responses amongst the students primed with the number 94. How is this apparent in the visualization? Compute the approximate values of a *summary statistic* we've seen in class to justify your answer.



c) The following graphic is created by the (incomplete) code snippet below.




```
ggplot(africa, aes(AAA, BBB)) +
  geom_CCC() +
  geom_DDD(method = "lm", se = FALSE) +
  labs(x = "Height (in inches)", y = "Number of countries in Africa guessed")
```

What precise code should be in place of `AAA`, `BBB`, `CCC`, and `DDD` in order to create this plot?

d) While an exploratory scatterplot of the relationship between y and `year` would be valid since `year` is numerical, why would a (vertical) boxplot with `year` on the x-axis also be acceptable *for this particular dataset*? Answer in one sentence.

4 Regression model using height

Continuing the previous `africa` question, say you run the following regression instead, using `height` instead of `priming` as the explanatory/predictor variable:

```
model_countries_height <- lm(how_many_countries ~ height, data = africa)
get_regression_table(model_countries_height)

## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   162.       86.9       1.86   0.07    -14.1   338.
## 2 height       -1.86      1.34     -1.39   0.174    -4.57    0.854
```

a) Interpret the `intercept` term in the `estimate` column of the regression table, both mathematically and practically speaking ("practically" meaning in context of the data).

b) Give the precise interpretation of the slope for `height` in the `estimate` column of the regression table.

c) Say we run the following code and present only the first row of the output (out of 41 rows), corresponding to the first student in `africa`. What are `XXX` and `YYY`? Your answers should be numerical values. Show your work.

```
get_regression_points(model_countries_height)
```

```
## # A tibble: 1 x 5
##      ID how_many_countries height how_many_countries_hat residual
##   <int>              <dbl>  <dbl> <chr>                  <chr>
## 1     1                 36     70 XXX                    YYY
```

d) Based on the regression model above, someone predicts that someone of height 54 inches will guess 62 countries. Why might this prediction inappropriate? Base your answer only on the various output of the analysis/model so far, and not prior knowledge or hypotheses you may have about the relationship between height and knowledge of the number of countries in Africa.

e) What would it mean for the relationship between height and the number of countries guessed if the slope for `height` in the table above were 0? Answer in practical and not mathematical terms ("practical" meaning in context of the data).

f) Do you think the observed slope for `height` of `-1.86` is *significantly* different from 0? Why? You will receive full credit for merely making a good faith attempt at answering. A "right answer" is not expected as you don't have the tools to answer this question . . . yet.