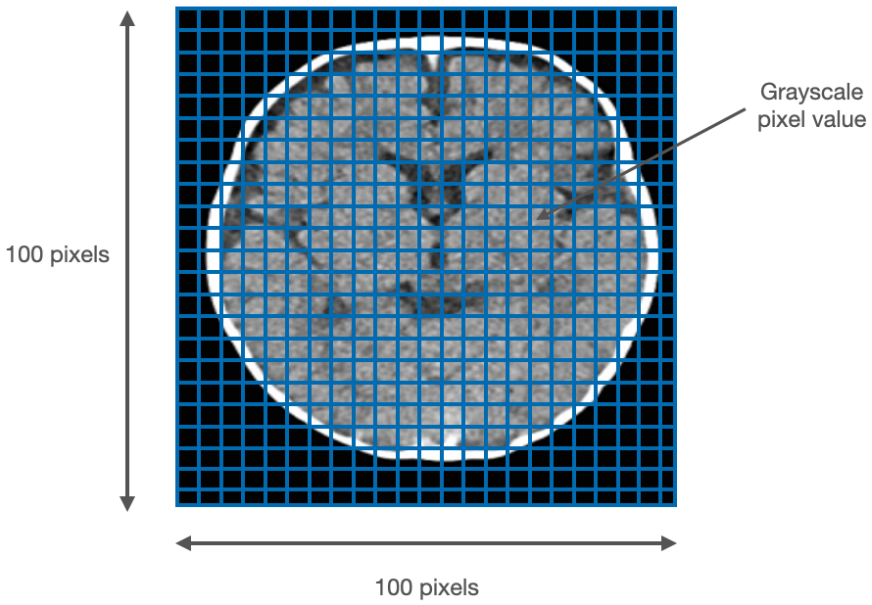
Quiz 1

You have 30 minutes to complete this 10 question quiz. The questions, a mix of multiple choice, fill-in-the-blank, and numeric answers, are weighted equally. You can consult any course materials or the internet. However, you cannot use R and you must complete the quiz individually.

1 Fill in the Blank 0.5 points

We have an electronic health record database with information on 1000 patients. For each patient, we have available a 2-dimensional CT image, which is a 100x100 array of pixels with each pixel represented by a continuous grayscale value between 0 and 1 (see below) as well as a binary stroke type response (ischemic or hemorrhagic). Each pixel is a feature.

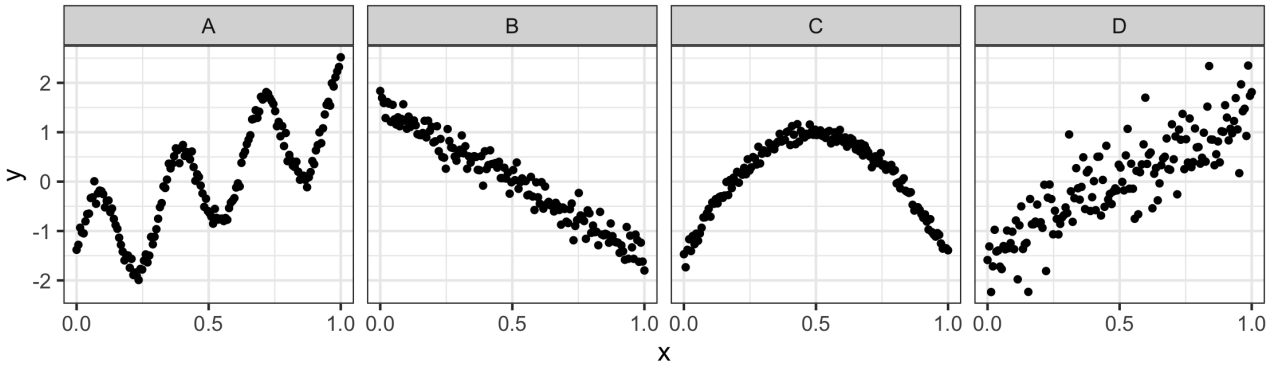


The data are represented using a table of the kind in Lecture 1. What are the dimensions of this table?

This table has 1000 rows and 10001 columns.

2 Multiple Choice 0.5 points

A simple linear regression is run for each of the four scatter plots below. Which leads to the highest R^2 value?



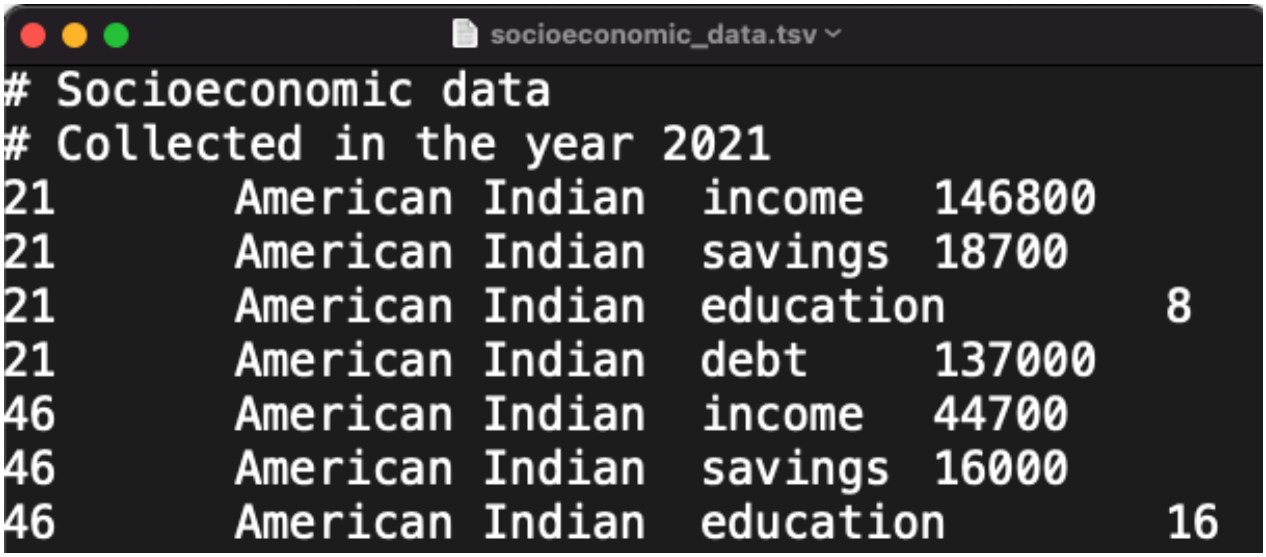
- A
- B
- C
- D

3 Fill in the Blank 0.5 points

Socioeconomic data

The next five questions concern a hypothetical socioeconomic dataset containing the age, race, and four economic indicators for a set of 20 individuals.

The socioeconomic data are stored in a .tsv file, the first few lines of which are shown below:



We read this file into R using read_tsv().

To avoid reading in the header, we can use the arguments skip = 2 or comment = "#".

4 Numeric 0.5 points

Continuing where question 3 left off, suppose we successfully avoid reading in the header. Furthermore, suppose we specify col_names = TRUE in the call to read_tsv(). How many rows of data will the resulting tibble contain?

79

5 Fill in the Blank 0.5 points

We realize that col_names = FALSE would be a better choice, and successfully read the tibble into R:

```
> socioeconomic_data
# A tibble: 80 × 4
     age race                  indicator  value
   <int> <fct>                 <fct>      <dbl>
 1    21 American Indian       income    146800
 2    21 American Indian       savings    18700
 3    21 American Indian       education      8
 4    21 American Indian       debt      137000
 5    46 American Indian       income     44700
 6    46 American Indian       savings    16000
 7    46 American Indian       education     16
 8    46 American Indian       debt       97800
 9    46 Pacific Islander      income    135100
10    46 Pacific Islander      savings    40800
# … with 70 more rows
```

To tidy this data, we apply the following pivot operation:

```
> socioeconomic_data %>%
      pivot_wider(names_from = indicator, values_from = value)
```

The resulting tibble contains [20] rows and

[6] columns.

6 Fill in the Blank 0.5 points

Instead of tidying, suppose we summarize the original tibble as follows:

```
> socioeconomic_data %>%
      group_by(indicator) %>%
      summarise(mean_value = mean(value))
```

The resulting tibble will contain 4 rows and 2 columns.

7 Fill in the Blank 0.5 points

Instead of tidying or summarizing, suppose we instead transform the original tibble using the following sequence of steps:

```
> socioeconomic_data %>%
      filter(indicator == "income") %>%
      mutate(high_income = value > 100000) %>%
      select(age, race, high_income)
```

The resulting tibble will contain 20 rows and 3 columns.

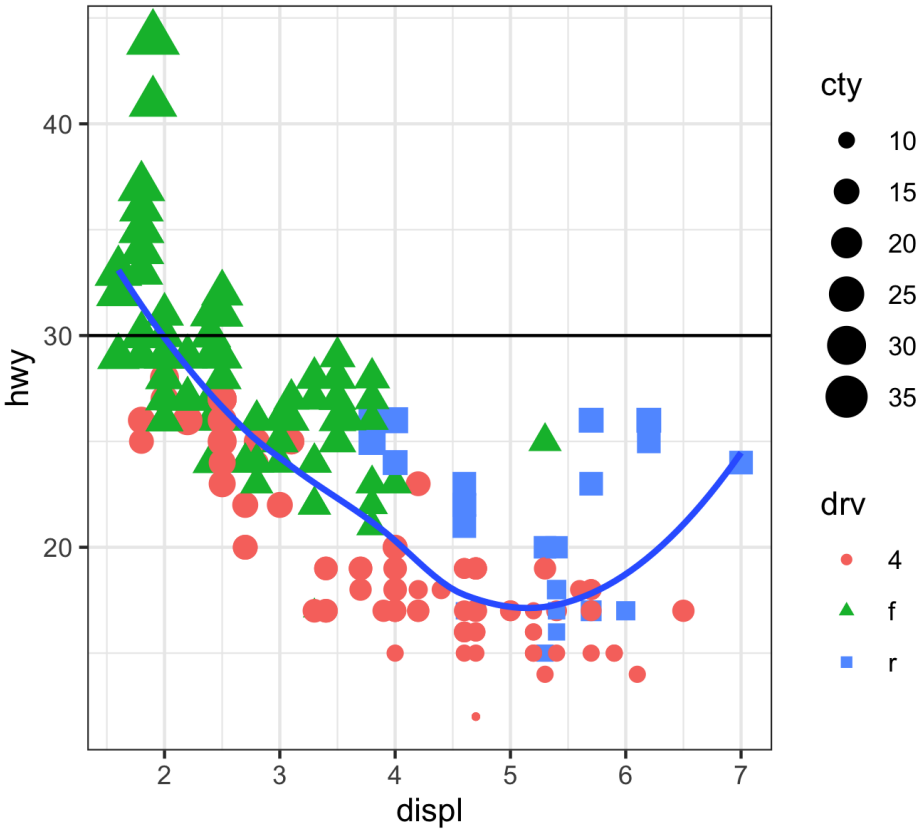
8 Numeric 0.5 points

mpg data

The next three questions concern the mpg data discussed in Lecture 4:

```
> mpg
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual(m6) 4        20    28 p     compact
# … with 224 more rows
```

Consider the ggplot below:



How many distinct geoms does this plot contain?

3

9 Fill in the Blank 0.5 points

Below is the beginning of the code used to produce the plot above. Fill in the blanks, using exactly one space before and after equal signs. Note that order does not matter; i.e. filling the blanks in any way that produces the correct plot will be counted as correct.

```
mpg %>% ggplot(aes(x = displ, y = hwy)) + geom_point(aes(
```

shape = drv , size = cty ,

color = drv)) + ...

10 Multiple Answer 0.5 points

Which of the following code chunks has the effect of filtering the mpg data frame to retain cars manufactured in Japan (i.e. manufactured by Honda, Subaru, Toyota, or Nissan)? Select all that apply.

```
# Option A
mpg %>%
  filter(manufacturer == "honda" |
         manufacturer == "subaru" |
         manufacturer == "toyota" |
         manufacturer == "nissan")

# Option B
mpg %>%
  filter(manufacturer == "honda" &
         manufacturer == "subaru" &
         manufacturer == "toyota" &
         manufacturer == "nissan")

# Option C
mpg %>%
  filter(manufacturer %in% c("honda", "subaru", "toyota", "nissan"))

# Option D
mpg %>%
  filter(manufacturer == "honda") %>%
  filter(manufacturer == "subaru") %>%
  filter(manufacturer == "toyota") %>%
  filter(manufacturer == "nissan")
```

- [x] A
- [ ] B
- [x] C
- [ ] D