Homework 3

MTH 3270

Tobias Boggess

2/13/2022

Homework 3

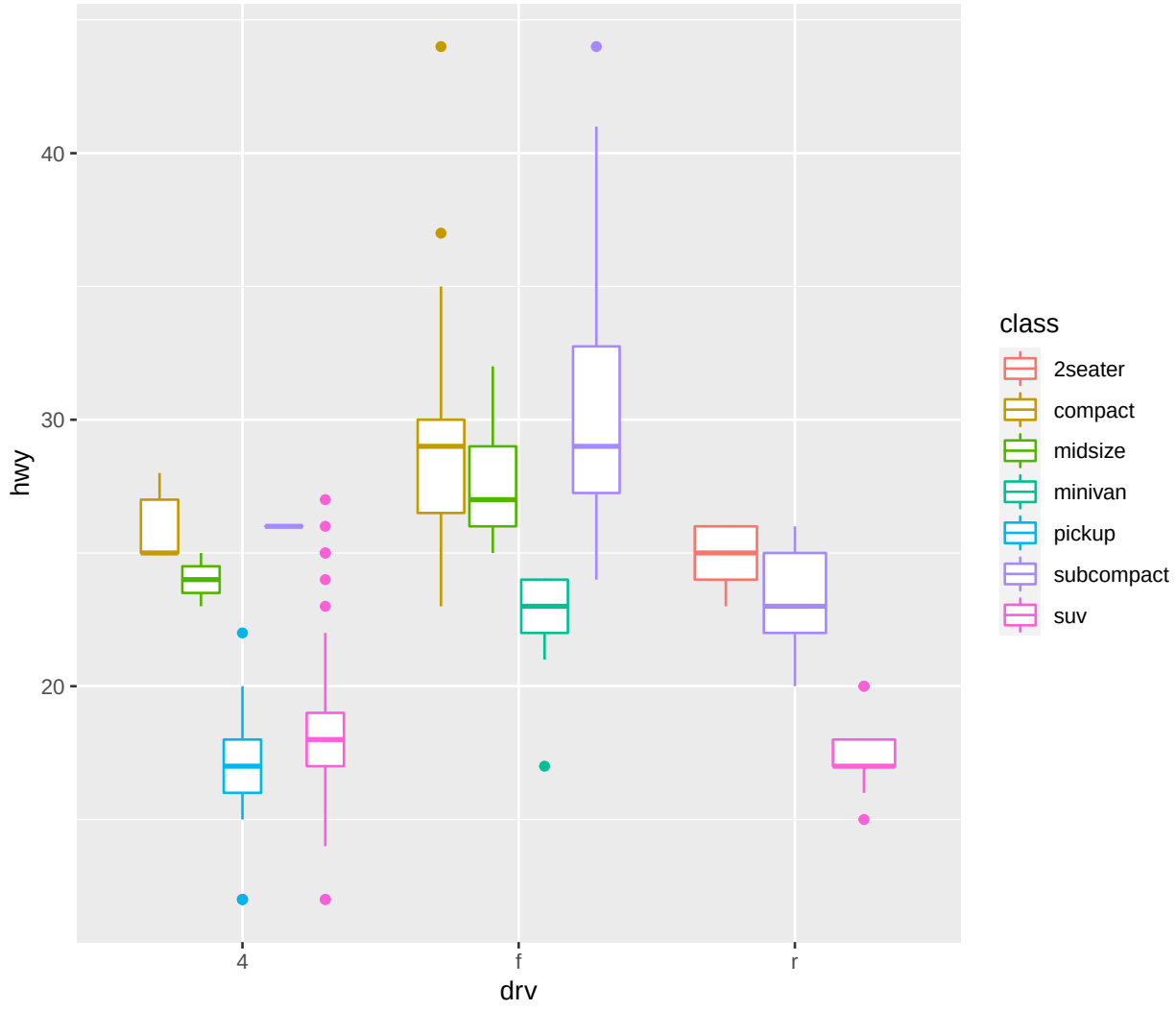
Problems on worksheet

Problem 1: Answer the questions given below.

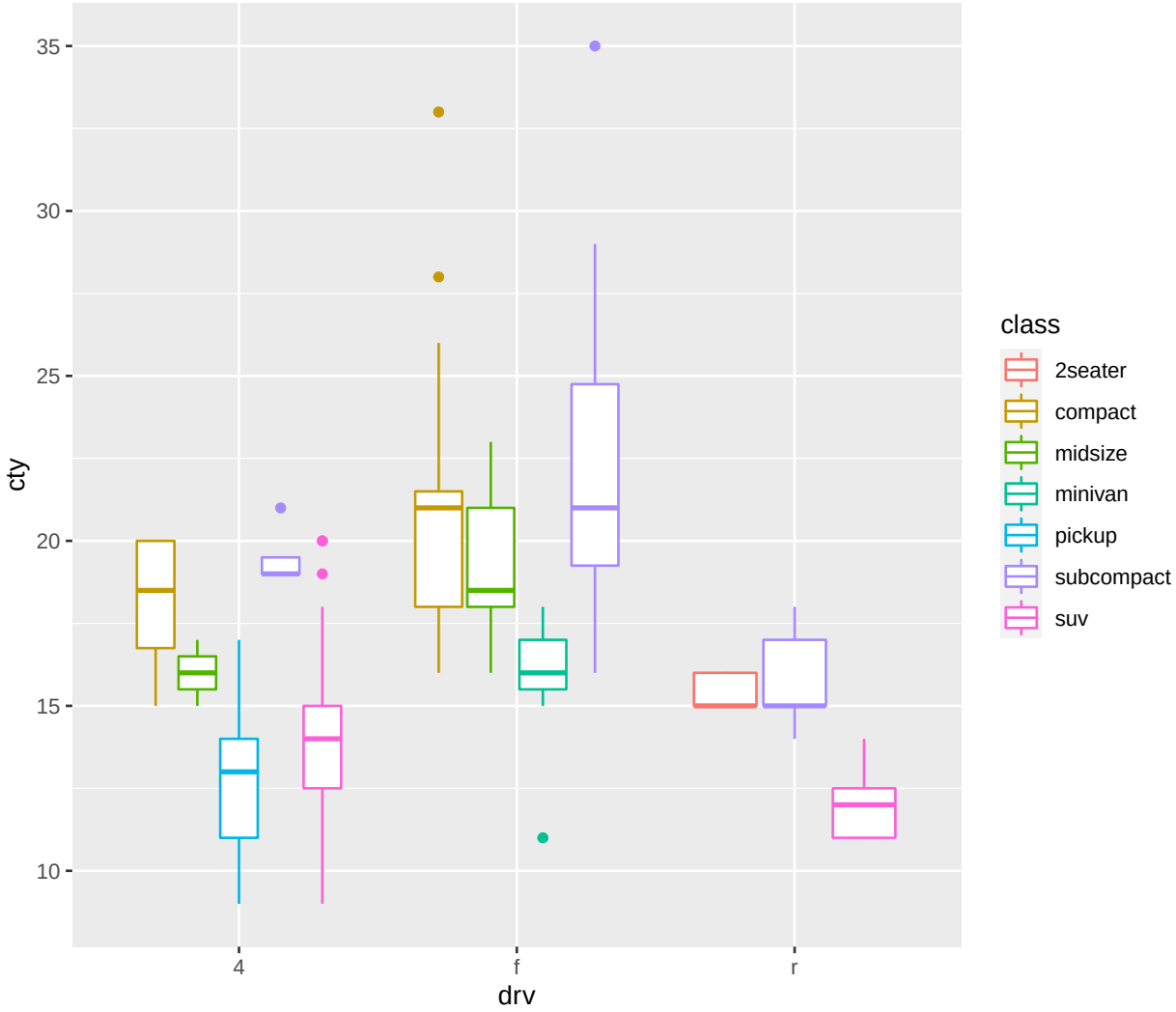
a) What's the default position adjustment according to the help page for geom_boxplot?
The default position for *geom_boxplot* is *"dodge2"* according to the help page.

b) Visualization of mpg data set using *geom_boxplot*
Graph:

```
library(ggplot2)
ggplot(data = mpg) +
  geom_boxplot(mapping = aes(x = drv, y = hwy, color = class))
```



```
ggplot(data = mpg) +
  geom_boxplot(mapping = aes(x = drv, y = cty, color = class))
```



Both of the graphs decided were based on the drive type, fuel economy, and the class type.

Problem 2: Use the data from the *nels88.txt* file.

a) Save the file and read it into *read.csv*
Code:

```
nelsFile <- file.choose()
nels <- read.csv(nelsFile, header = TRUE, sep = "")
head(nels)
```

```
##       id sch_id heldback         schtype                 race    ses female
## 1 175507   1755       NO CATHOLIC SCHOOL   WHITE NOT HISPANIC  0.912      1
## 2 175517   1755       NO CATHOLIC SCHOOL   BLACK NOT HISPANIC  0.761      1
## 3 175521   1755       NO CATHOLIC SCHOOL ASIAN/PACIFIC ISLNDR  0.786      1
## 4 175528   1755       NO CATHOLIC SCHOOL   WHITE NOT HISPANIC -0.019      1
## 5 175544   1755      YES CATHOLIC SCHOOL             HISPANIC  0.089      0
## 6 175550   1755       NO CATHOLIC SCHOOL   WHITE NOT HISPANIC -0.014      1
##   minority asian hispanic black white native catholic private bymath f1math
## 1        0     0        0     0     1      0        1       0  37.40     NA
## 2        1     0        0     1     0      0        1       0  46.73     NA
## 3        1     1        0     0     0      0        1       0  36.34     NA
## 4        0     0        0     0     1      0        1       0  49.16     NA
## 5        1     0        1     0     0      0        1       0  45.10     NA
## 6        0     0        0     0     1      0        1       0  38.44     NA
##   f2math f2dropout
## 1     NA      <NA>
## 2     NA      <NA>
## 3     NA      <NA>
## 4     NA      <NA>
## 5     NA      <NA>
## 6     NA      <NA>
```

```
str(nels)
```

```
## 'data.frame':    6170 obs. of  19 variables:
##  $ id       : int  175507 175517 175521 175528 175544 175550 175551 175558 175560 175568 ...
##  $ sch_id   : int  1755 1755 1755 1755 1755 1755 1755 1755 1755 1755 ...
##  $ heldback : chr  "NO" "NO" "NO" "NO" ...
##  $ schtype  : chr  "CATHOLIC SCHOOL" "CATHOLIC SCHOOL" "CATHOLIC SCHOOL" "CATHOLIC SCHOOL" ...
##  $ race     : chr  "WHITE NOT HISPANIC" "BLACK NOT HISPANIC" "ASIAN/PACIFIC ISLNDR" "WHITE NOT HISPA
##  $ ses      : num  0.912 0.761 0.786 -0.019 0.089 ...
##  $ female   : int  1 1 1 1 0 1 1 0 1 1 ...
##  $ minority : int  0 1 1 0 1 0 1 0 1 1 ...
##  $ asian    : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ hispanic : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ black    : int  0 1 0 0 0 0 1 0 0 1 ...
##  $ white    : int  1 0 0 1 0 1 0 1 0 0 ...
##  $ native   : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ catholic : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ private  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bymath   : num  37.4 46.7 36.3 49.2 45.1 ...
##  $ f1math   : num  NA NA NA NA NA ...
##  $ f2math   : num  NA NA NA NA NA ...
##  $ f2dropout: chr  NA NA NA NA ...
```

```
names(nels)
```

```
##  [1] "id"        "sch_id"    "heldback"  "schtype"   "race"      "ses"
##  [7] "female"    "minority"  "asian"     "hispanic"  "black"     "white"
## [13] "native"    "catholic"  "private"   "bymath"    "f1math"    "f2math"
## [19] "f2dropout"
```

```
nrow(nels)
```

```
## [1] 6170
```

```
ncol(nels)
```

```
## [1] 19
```

```
View(nels)
is.numeric(nels)
```

```
## [1] FALSE
```

There are 6,170 rows in the *nels88.txt* data and 19 columns. Additionally, 15 of the columns are numerical and the other four columns are categorical.

b) Number of "complete" rows (i.e., no values showing as NA)
Code:

```
sum(complete.cases(nels))
```

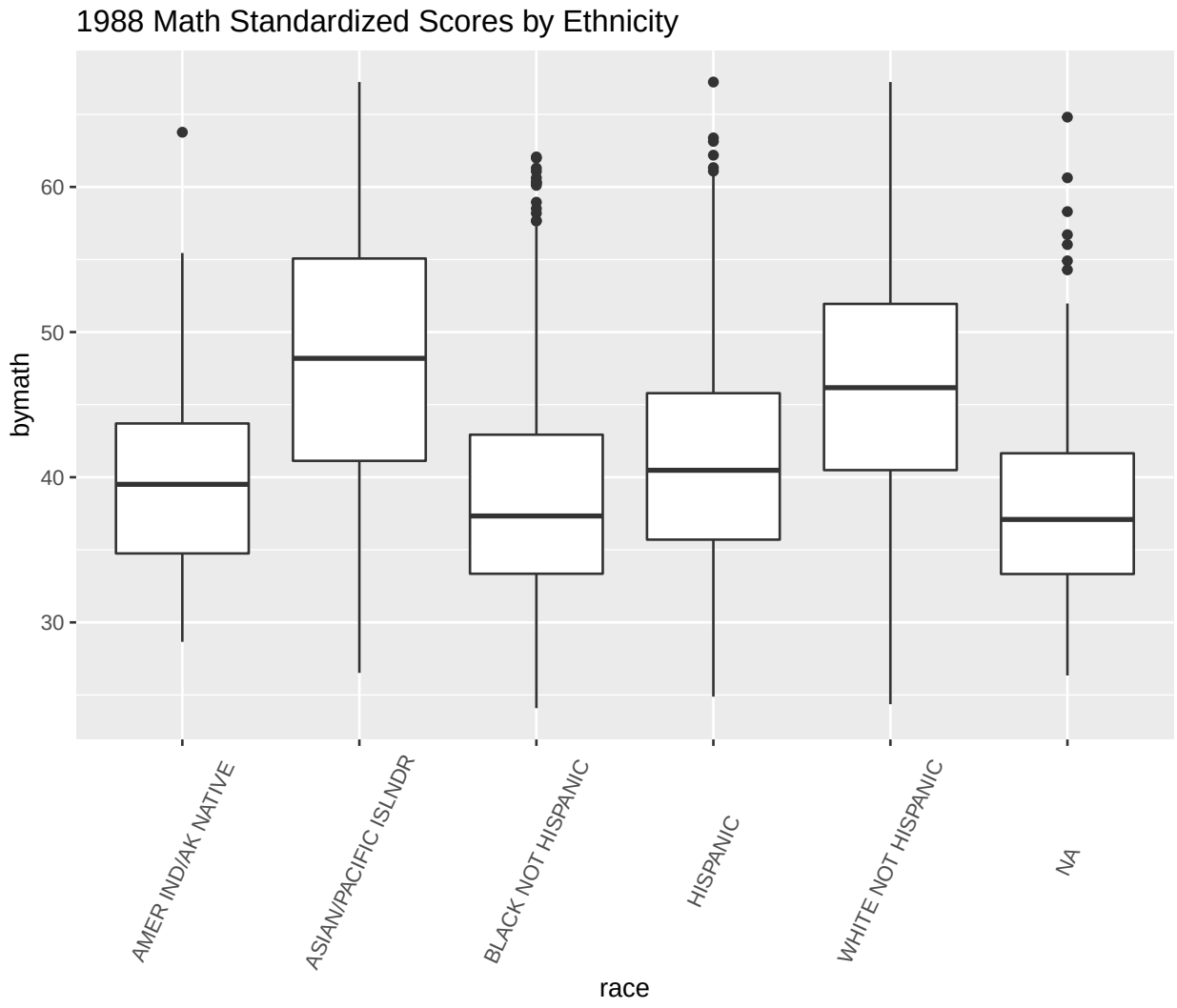
```
## [1] 2660
```

There are 2660 rows that are complete.

c) Create three different graphs, one of which needs at least three different variables. Show the R commands.
Code:

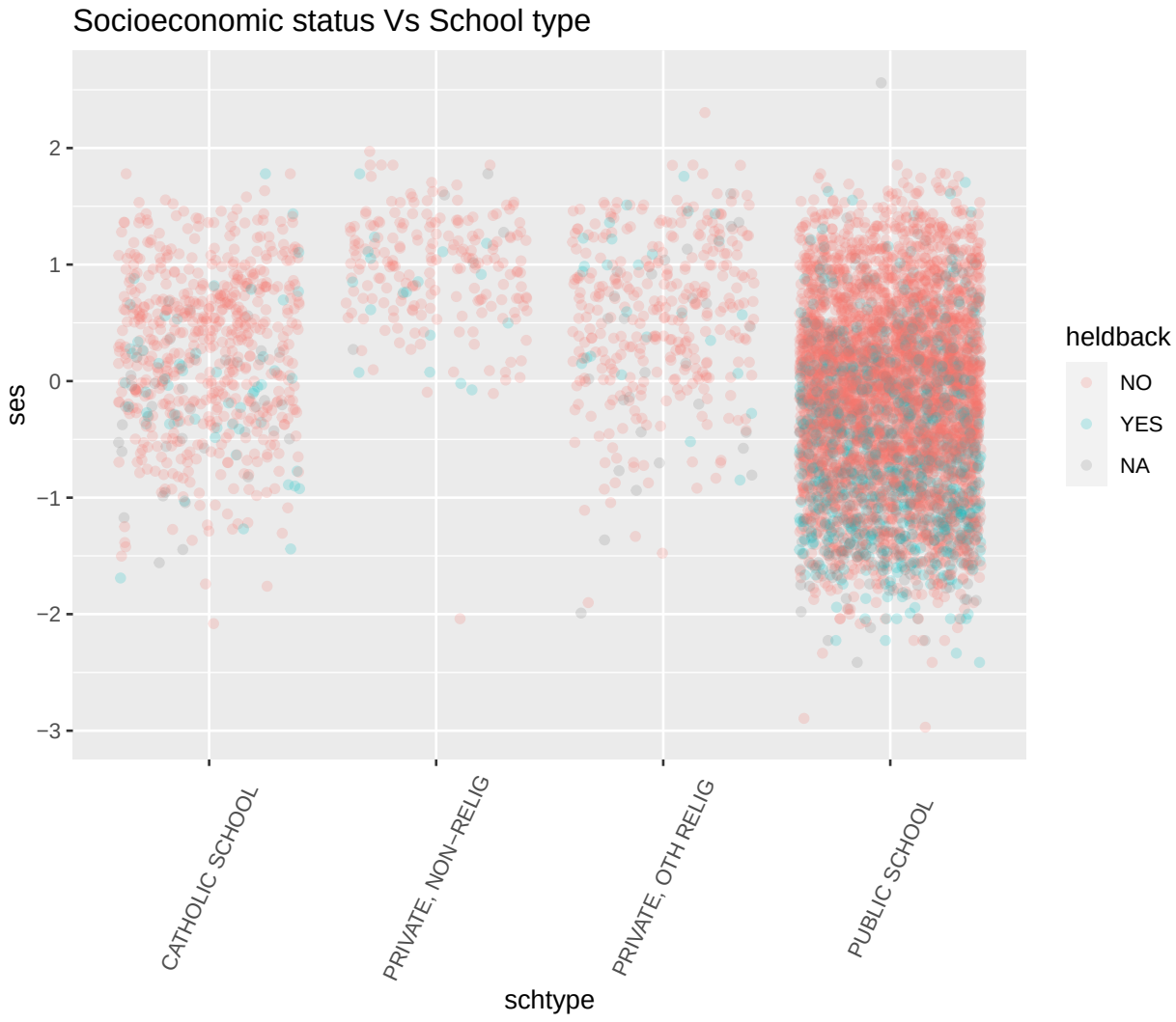
```
ggplot(nels) +
  geom_boxplot(mapping = aes(x = race, y = bymath, color = female)) +
  labs(title = "1988 Math Standardized Scores by Ethnicity") +
  theme(axis.text.x = element_text(angle = 65, vjust = 0.5))
```

```
## Warning: Removed 227 rows containing non-finite values (stat_boxplot).
```



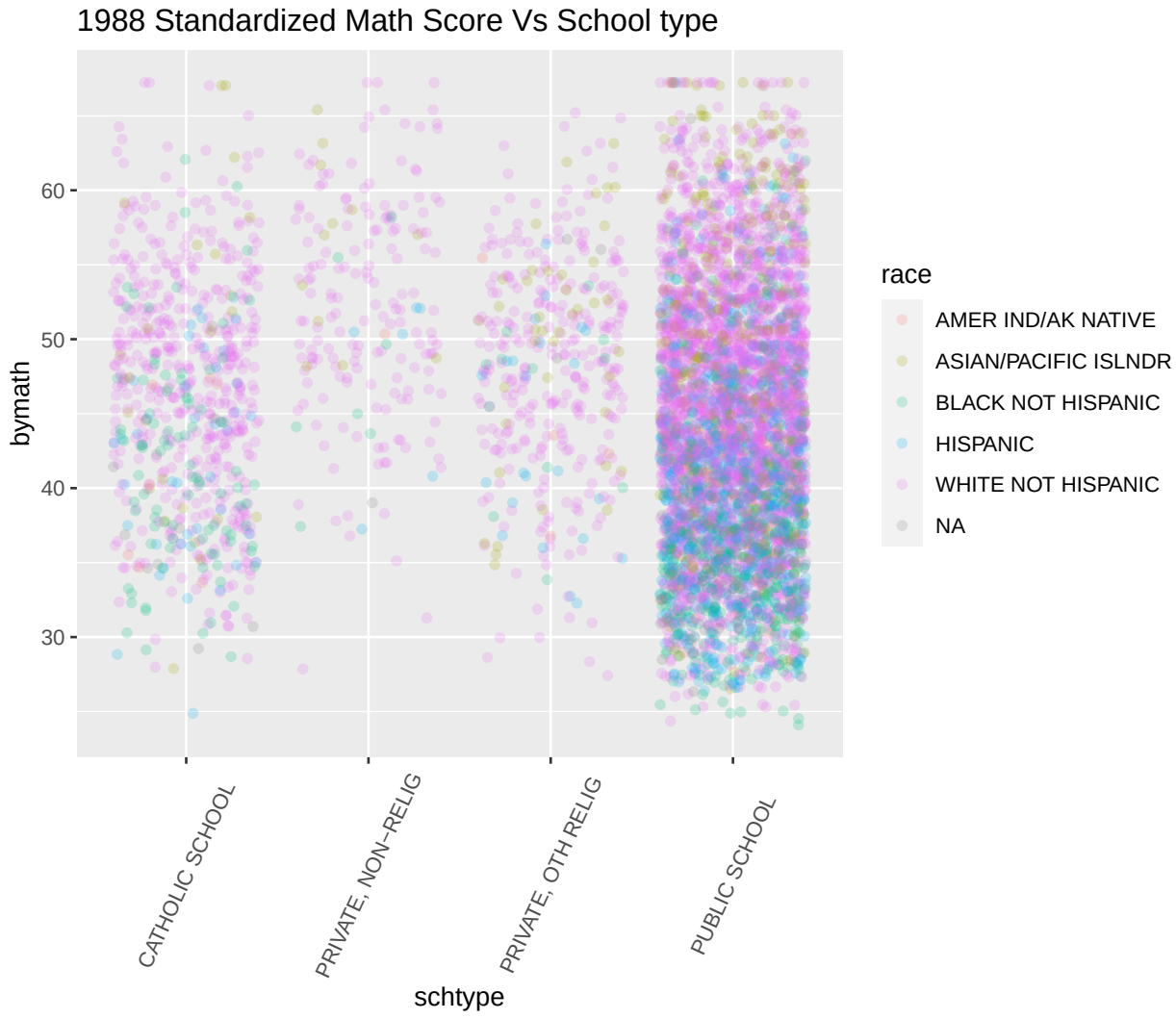
```
ggplot(nels) +
  geom_point(
    mapping = aes(
      x = schtype,
      y = ses,
      color = heldback,
    ),
    alpha = 0.2,
    position = "jitter"
  ) +
  labs(title = "Socioeconomic status Vs School type") +
  theme(axis.text.x = element_text(angle = 65, vjust = 0.5))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



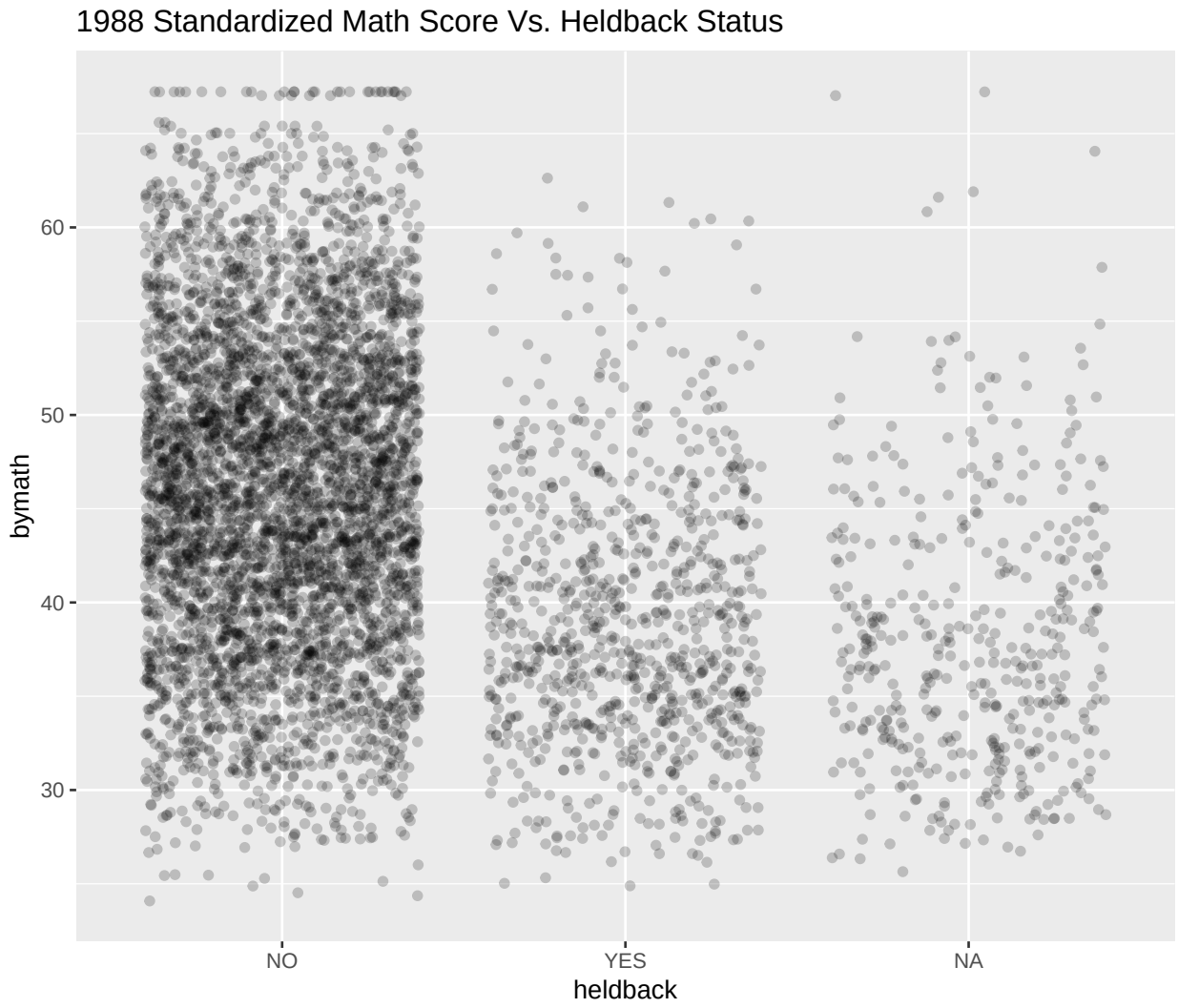
```
ggplot(nels) +
  geom_point(
    mapping = aes(
      x = schtype,
      y = bymath,
      color = race,
    ),
    alpha = 0.2,
    position = "jitter"
  ) +
  labs(title = "1988 Standardized Math Score Vs School type") +
  theme(axis.text.x = element_text(angle = 65, vjust = 0.5))
```

```
## Warning: Removed 227 rows containing missing values (geom_point).
```



```
ggplot(nels) +
  geom_point(mapping = aes(x = heldback, y = bymath), alpha = 0.2, position = "jitter") +
  labs(title = "1988 Standardized Math Score Vs. Heldback Status")
```

```
## Warning: Removed 227 rows containing missing values (geom_point).
```

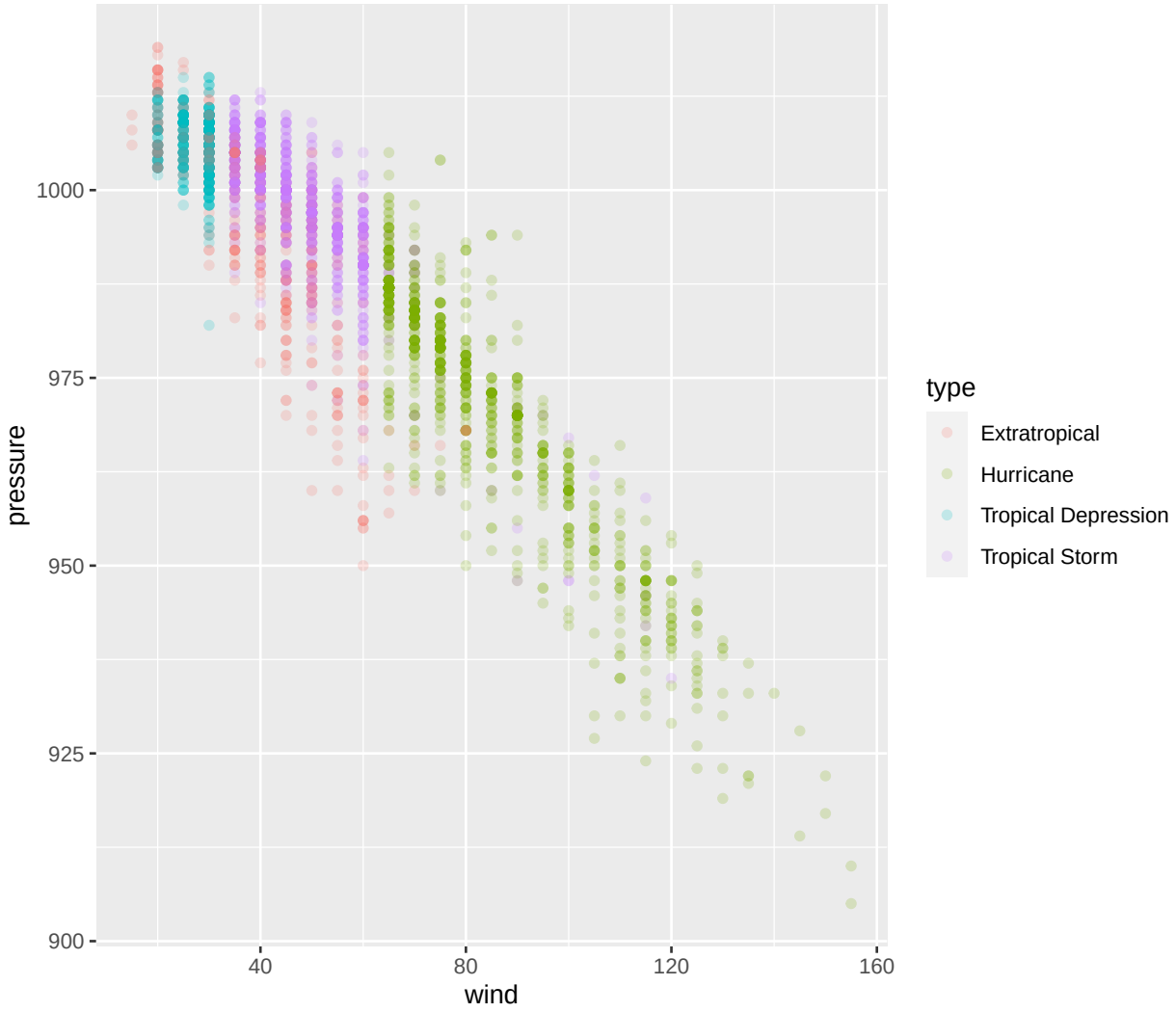


Book Problems

Problem 2: Create a scatterplot between wind and pressure, with color being used to distinguish the type of storm.

Code:

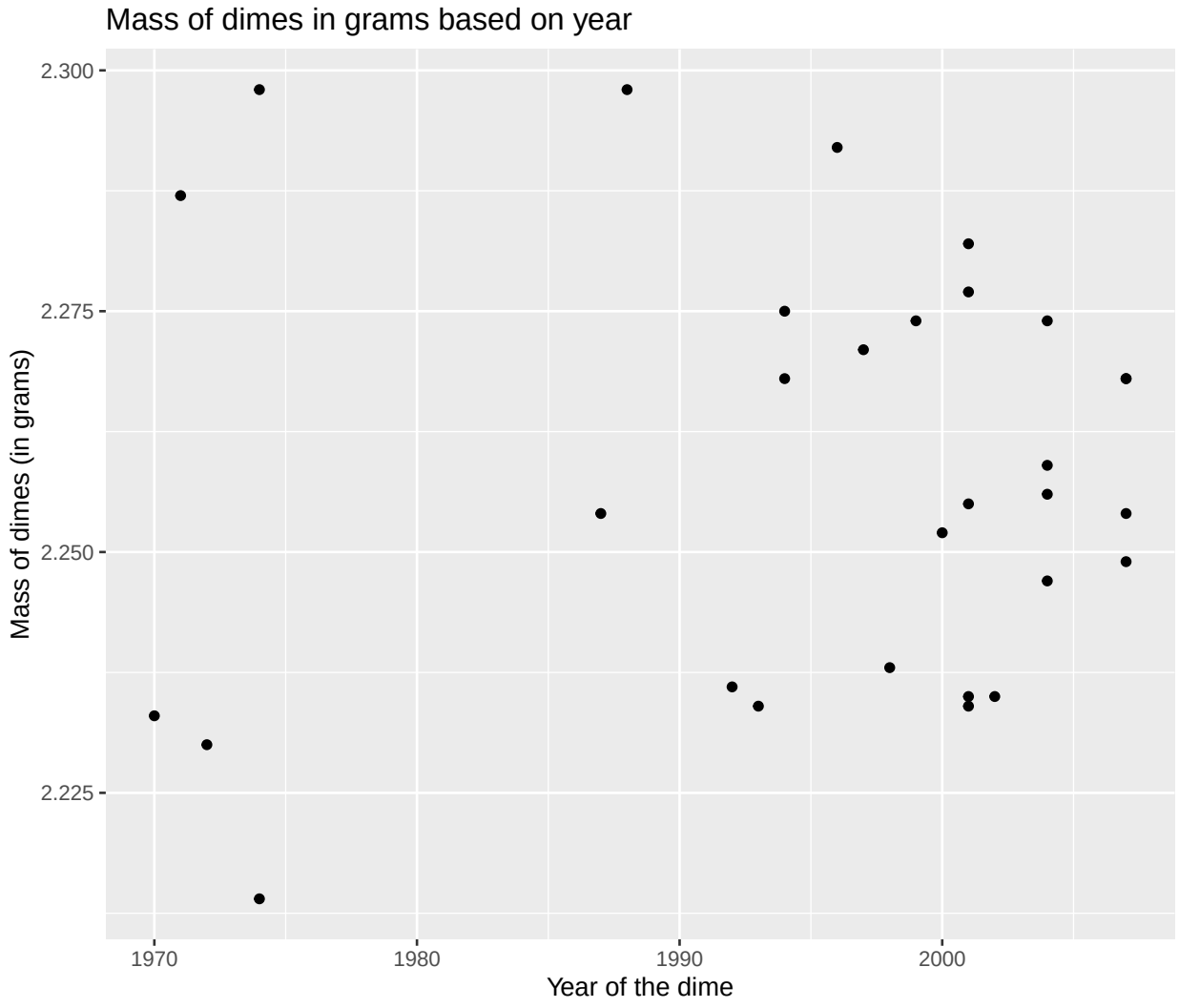
```
library(ggplot2)
library(nasaweather)
ggplot(data = storms) +
  geom_point(
    mapping = aes(x = wind, y = pressure, color = type),
    alpha = 0.2
  )
```



Problem 3: Do the following.

a) Create an informative and meaningful data graphic
Code:

```
library(ggplot2)
library(mosaicData)
ggplot(data = Dimes) +
  geom_point(mapping = aes(x = year, y = mass)) +
  labs(title = "Mass of dimes in grams based on year",
       y = "Mass of dimes (in grams)",
       x = "Year of the dime"
       )
```

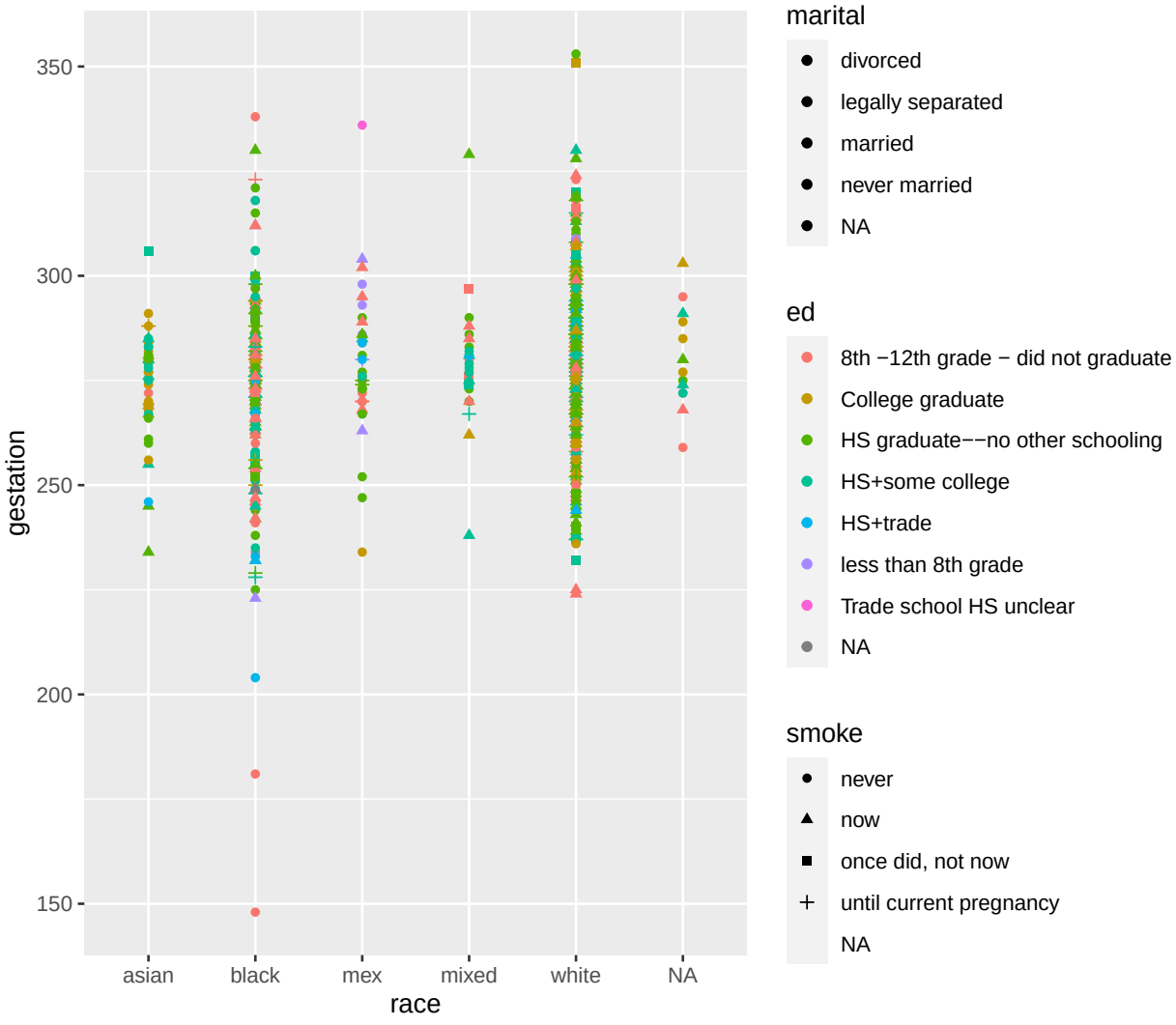


b) Identify each of the visual cues that you are using, and describe how they are related to each variable.
In the graph from *part a*, I used position to demonstrate the weights of the dimes according to year.

c) Create a data graphic with at least five variables (either quantitative or categorical).
Code:

```
ggplot(data = Gestation) +
  geom_point(mapping = aes(
    x = race,
    y = gestation,
    color = ed,
    fill = marital,
    shape = smoke
  ))
```

```
## Warning: Removed 23 rows containing missing values (geom_point).
```



Problem 4: Using ggplot2, create a data graphic that displays the average temperature over each 10-minute interval (temperature) as a function of time (when).

Code:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

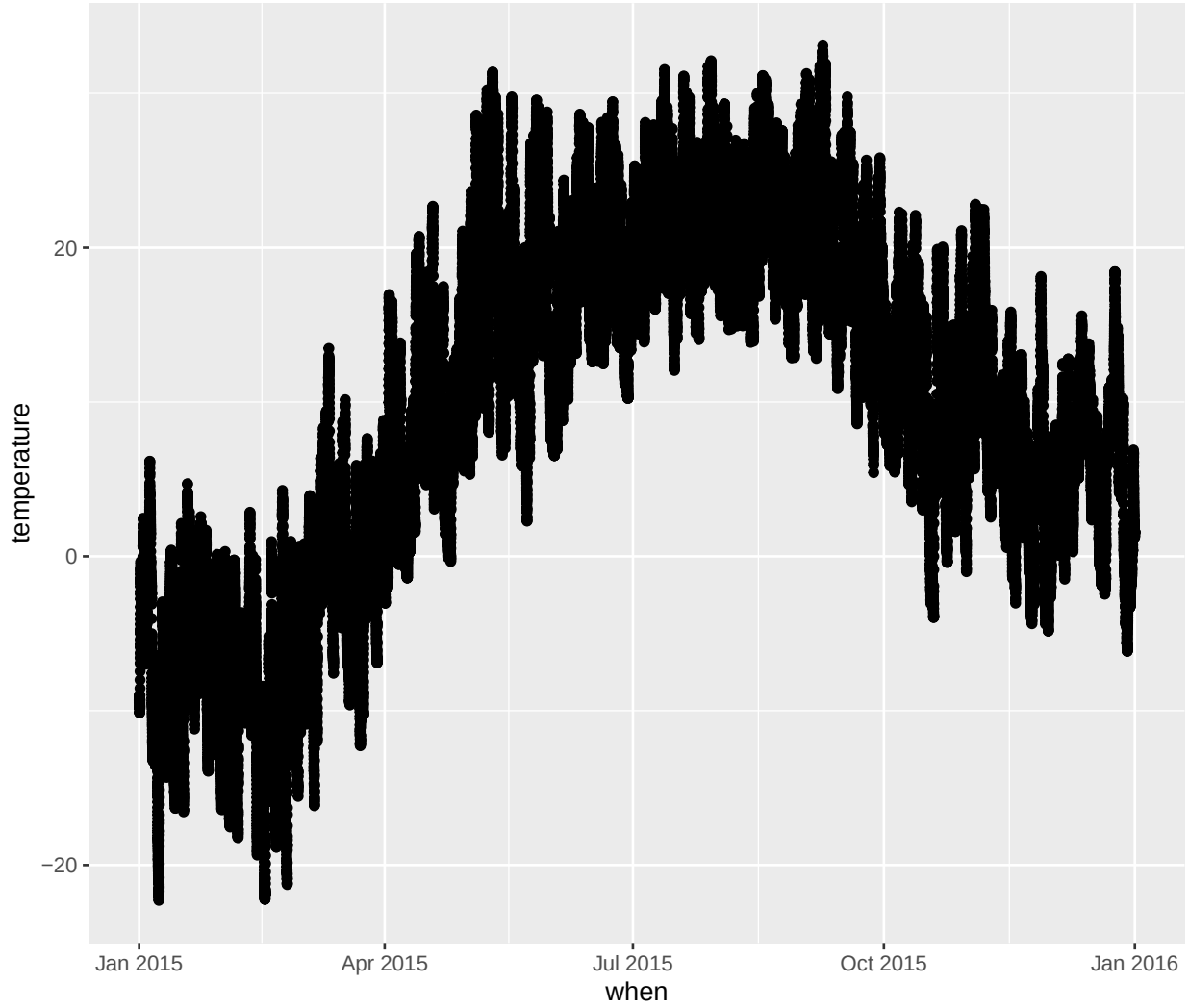
```
library(macleish)
```

```
## Loading required package: etl
```

```
glimpse(whately_2015)
```

```
## Rows: 52,560
## Columns: 8
## $ when            <dttm> 2015-01-01 00:00:00, 2015-01-01 00:10:00, 2015-01-01 ~
## $ temperature     <dbl> -9.32, -9.46, -9.44, -9.30, -9.32, -9.34, -9.30, -9.10~
## $ wind_speed      <dbl> 1.399, 1.506, 1.620, 1.141, 1.223, 1.090, 1.168, 1.307~
## $ wind_dir        <dbl> 225.4, 248.2, 258.3, 243.8, 238.4, 241.7, 242.3, 244.2~
## $ rel_humidity    <dbl> 54.55, 55.38, 56.18, 56.41, 56.87, 57.25, 57.71, 58.16~
## $ pressure        <int> 985, 985, 985, 985, 984, 984, 984, 984, 984, 984, 984,~
## $ solar_radiation <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ rainfall        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

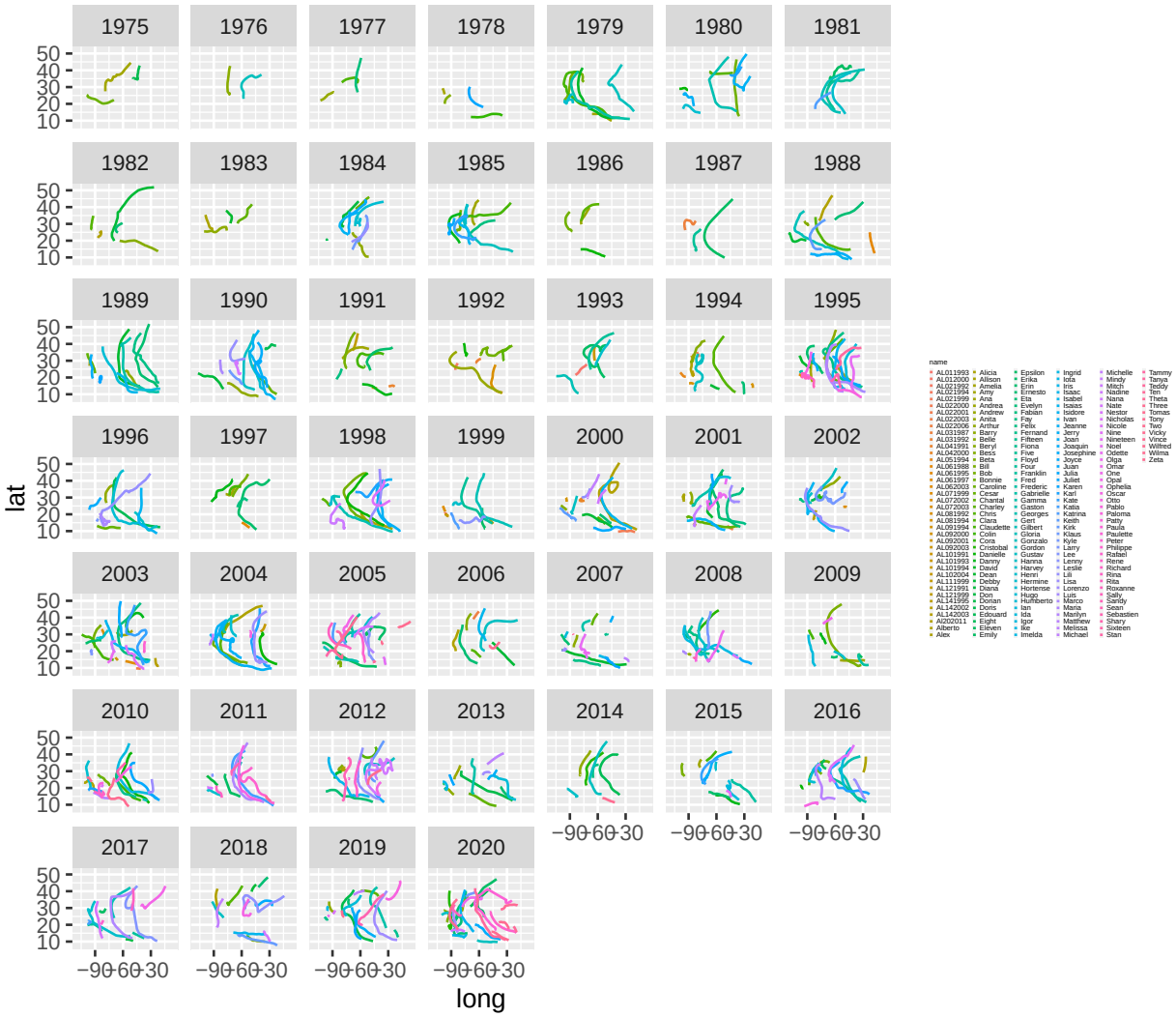
```
ggplot(data = whately_2015) +
  geom_point(mapping = aes(x = when, y = temperature))
```



Problem 8: Using data from the nasaweather package, use the geom_path function to plot the path of each tropical storm in the storms data table. Use color to distinguish the storms from one another, and use faceting to plot each year in its own panel.

Code:

```
library(nasaweather)
ggplot(data = storms,
       mapping = aes(x = long, y = lat, color = name)) +
  geom_path() +
  facet_wrap(facets = ~ year) +
  theme(
    legend.position = "right",
    legend.key.height = unit(0.001, 'cm'),
    legend.key.width = unit(0.001, 'cm'),
    legend.title = element_text(size = 3),
    legend.text = element_text(size = 3)
  ) +
  guides(col = guide_legend(nrow = 40))
```


1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
lat
long
50
40
30
20
10
−90
−60
−30
name

Problem 9: Do the following with the penguins data set.

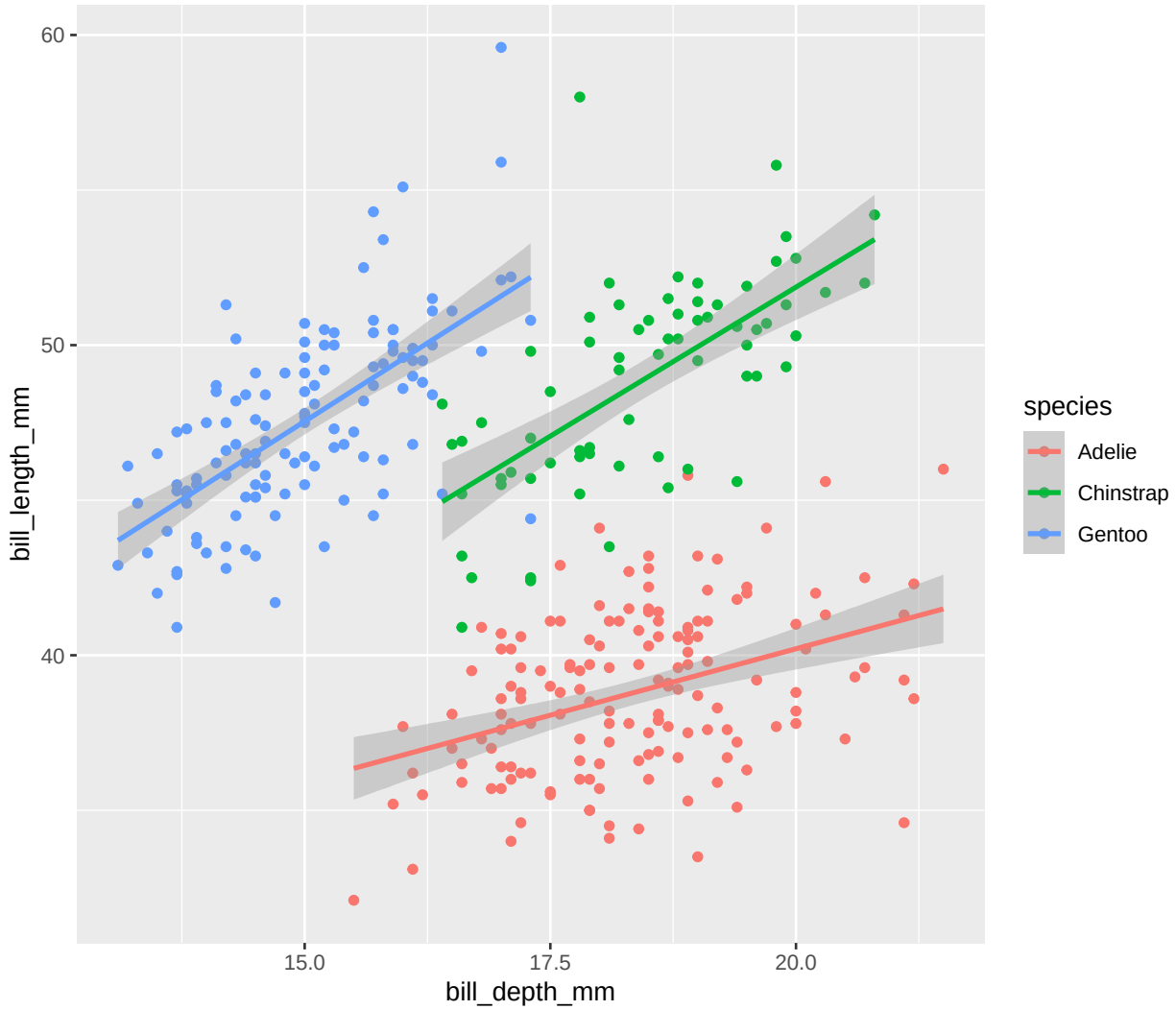
a) What do you observe about the association of bill depth and bill length?
Code:

```
library(palmerpenguins)
ggplot(
      data = penguins,
      mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)
      ) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```



As the bill depth increases on each species, the bill length will also increase in size.

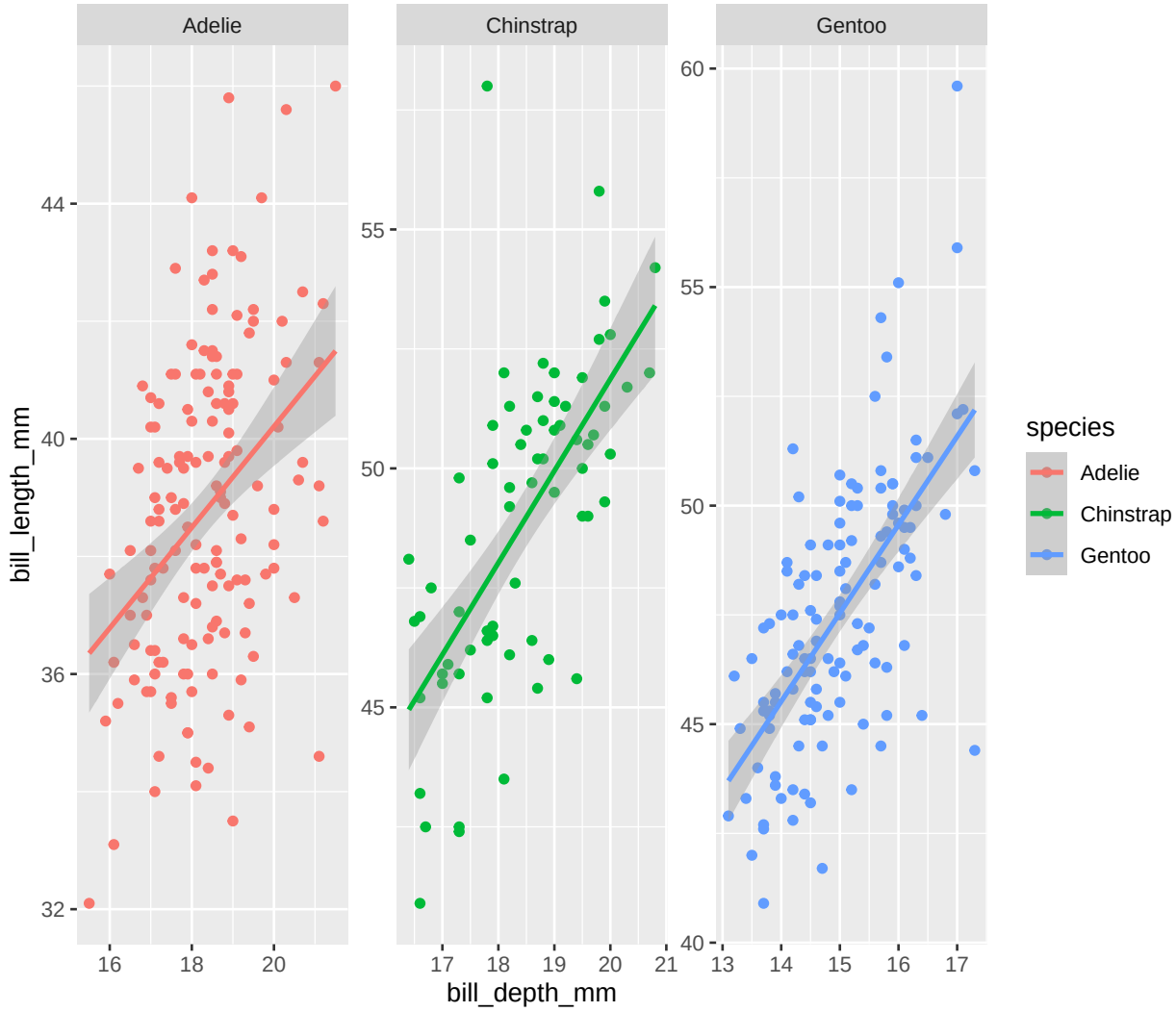
b) How would you summarize the association between bill depth and bill length.
Code:

```
ggplot(data = penguins, mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(facets = ~ species, scale = "free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



The relationship between bill depth and bill length seems linear and proportional. As the bill depth increases, the bill length increases.