L03 ggplot II

Data Visualization (STAT 302)

Anna Wagman

Contents

Overview

The goal of this lab is to continue the process of unlocking the power of `ggplot2` through constructing and experimenting with a few basic plots.

Datasets

We'll be using data from the `BA_degrees.rda` and `dow_jones_industrial.rda` datasets which are already in the `/data` subdirectory in our **data_vis_labs** project. Below is a description of the variables contained in each dataset.

`BA_degrees.rda`

- `field` - field of study
- `year_str` - academic year (e.g. 1970-71)
- `year` - closing year of academic year
- `count` - number of degrees conferred within a field for the year
- `perc` - field's percentage of degrees conferred for the year

`dow_jones_industrial.rda`

- `date` - date
- `open` - Dow Jones Industrial Average at open
- `high` - Day's high for the Dow Jones Industrial Average
- `low` - Day's low for the Dow Jones Industrial Average
- `close` - Dow Jones Industrial Average at close
- `volume` - number of trades for the day

We'll also be using a subset of the BRFSS (Behavioral Risk Factor Surveillance System) survey collected annually by the Centers for Disease Control and Prevention (CDC). The data can be found in the provided `cdc.txt` file — place this file in your `/data` subdirectory. The dataset contains 20,000 complete observations/records of 9 variables/fields, described below.

- `genhlth` - How would you rate your general health? (excellent, very good, good, fair, poor)
- `exerany` - Have you exercised in the past month? (`1 = yes`, `0 = no`)
- `hlthplan` - Do you have some form of health coverage? (`1 = yes`, `0 = no`)
- `smoke100` - Have you smoked at least 100 cigarettes in your life time? (`1 = yes`, `0 = no`)
- `height` - height in inches
- `weight` - weight in pounds

- `wtdesire` - weight desired in pounds
- `age` - in years
- `gender` - `m` for males and `f` for females

```
# Load package(s)
library(tidyverse)
library(skimr)
load(file = "data/BA_degrees.rda")
```

Exercises

Complete the following exercises.

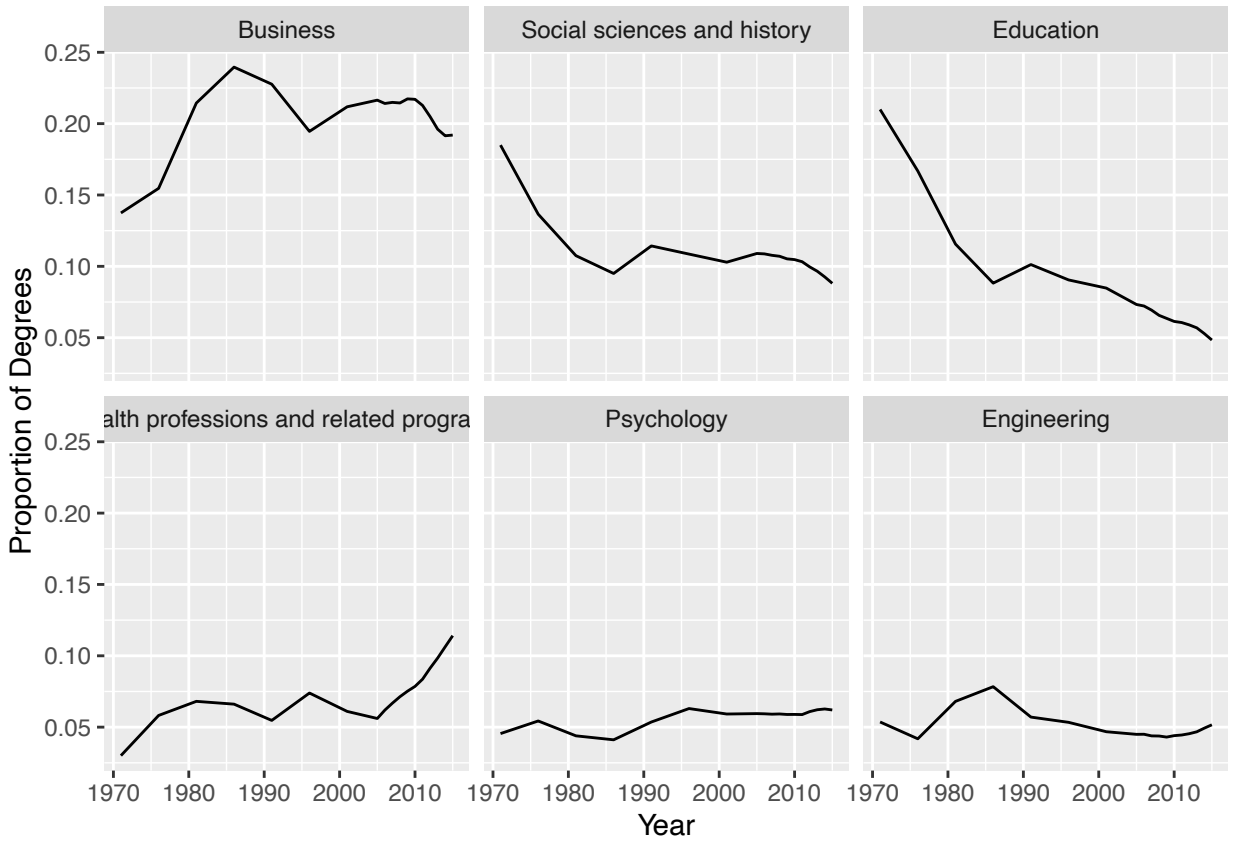
Exercise 1

Using `BA_degrees` dataset, recreate the following graphics as precisely as possible.

```
# Wrangling for plotting
ba_dat <- BA_degrees %>%
  # mean % per field
  group_by(field) %>%
  mutate(mean_perc = mean(perc)) %>%
  # Only fields with mean >= 5%
  filter(mean_perc >= 0.05) %>%
  # Organizing for plotting
  arrange(desc(mean_perc), year) %>%
  ungroup() %>%
  mutate(field = fct_inorder(field))
```

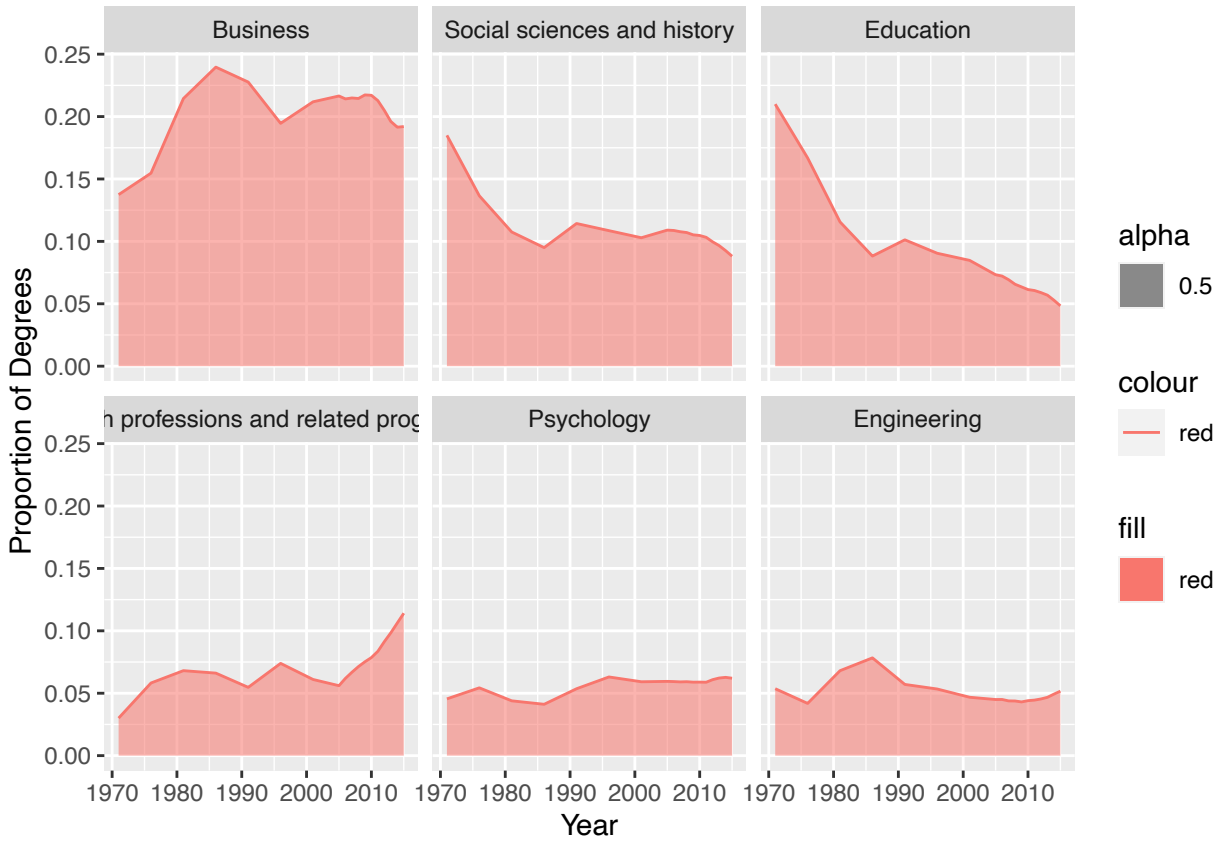
```
ba_dat <- BA_degrees %>%
  # mean % per field
  group_by(field) %>%
  mutate(mean_perc = mean(perc)) %>%
  # Only fields with mean >= 5%
  filter(mean_perc >= 0.05) %>%
  # Organizing for plotting
  arrange(desc(mean_perc), year) %>%
  ungroup() %>%
  mutate(field = fct_inorder(field))


ggplot(data = ba_dat, aes(year, perc)) +
  geom_line() +
  facet_wrap(~field) +
  labs(
    x = "Year",
    y = "Proportion of Degrees"
  )
```



Plot 1

```
ggplot(ba_dat, aes(year, perc)) +
  geom_area(aes(fill = "red", alpha = .5)) +
  facet_wrap(~field) + geom_line(aes(color = "red")) +
  labs(
    x = "Year",
    y = "Proportion of Degrees"
  )
```



Plot 2

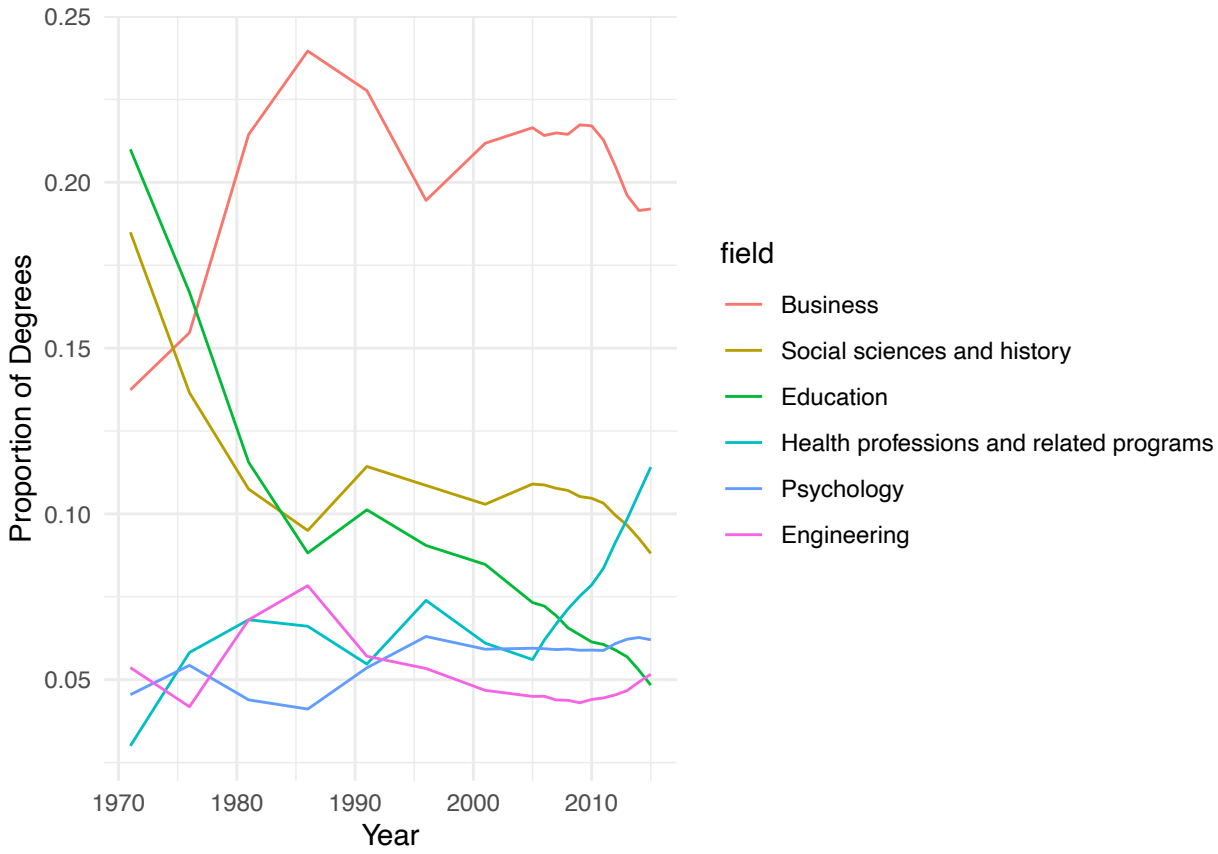
Hints:

- Transparency is 0.5
- Color used is`"red"`

```
ggplot(
  data = ba_dat,
  mapping = aes(x = year, y = perc, color = field)
) +
  geom_line() +

 labs(
  x = "Year",
  y = "Proportion of Degrees",


) +
  theme_minimal()
```



Plot 3

Exercise 2

Using `dow_jones_industrial` dataset, recreate the following graphics as precisely as possible. *Hint:* Used `close`.

```
load(file = "data/dow_jones_industrial.rda")
library(skimr)
library(tidyverse)
library(lubridate)

# Restrict data to useful range
djia_date_range <- dow_jones_industrial %>%
  filter(date >= ymd("2008/12/31") & date <= ymd("2010/01/10"))
```

```
load(file = "data/dow_jones_industrial.rda")
library(skimr)
library(tidyverse)
library(lubridate)
```

Plot 1

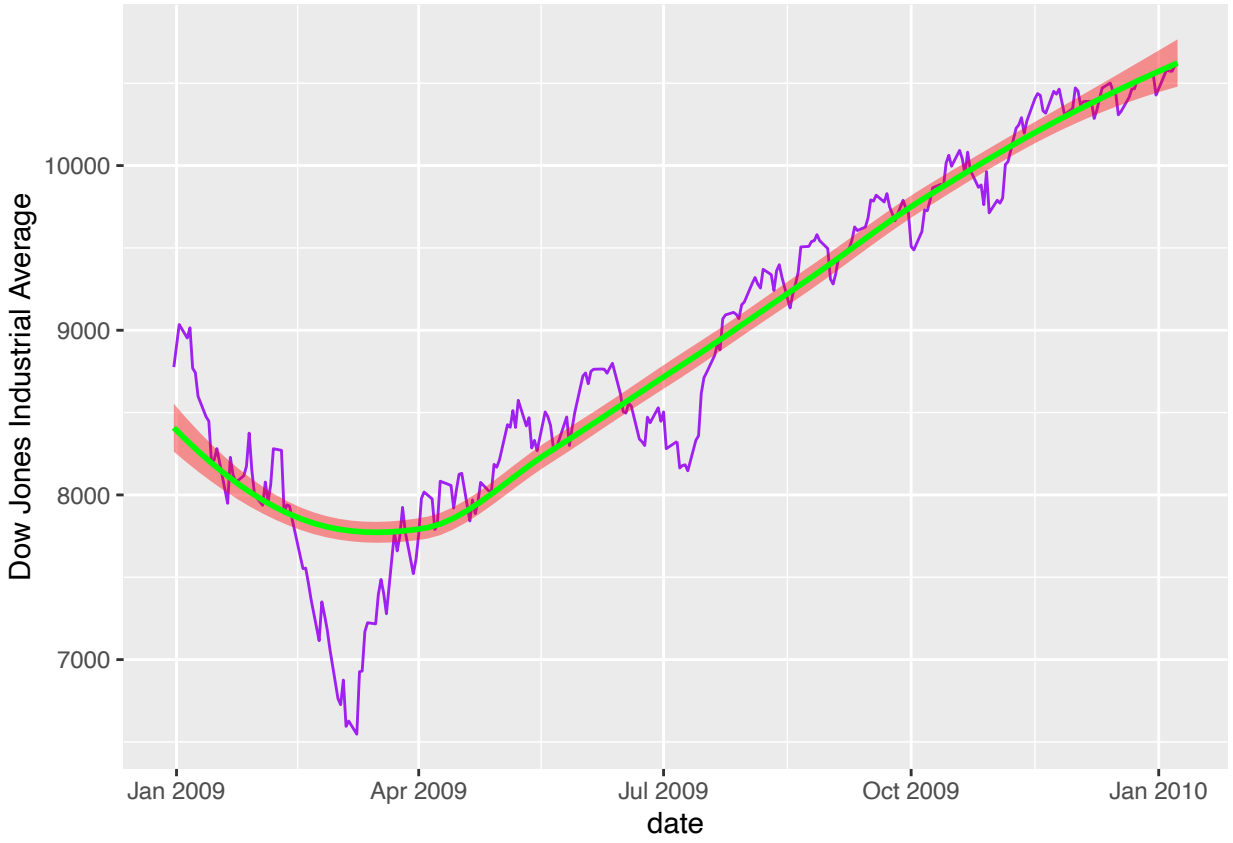
```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
djia_date_range <- dow_jones_industrial %>%
  filter(date >= ymd("2008/12/31") & date <= ymd("2010/01/10"))

ggplot(
  data = djia_date_range,
  mapping = aes(x = date, y = close)
) +
  geom_line(color = "purple") + geom_smooth(color = "green", fill = "red") +
  labs(
    y = "Dow Jones Industrial Average"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Hints:

- Colors used `"red"`, `"purple"`, & `"green"`

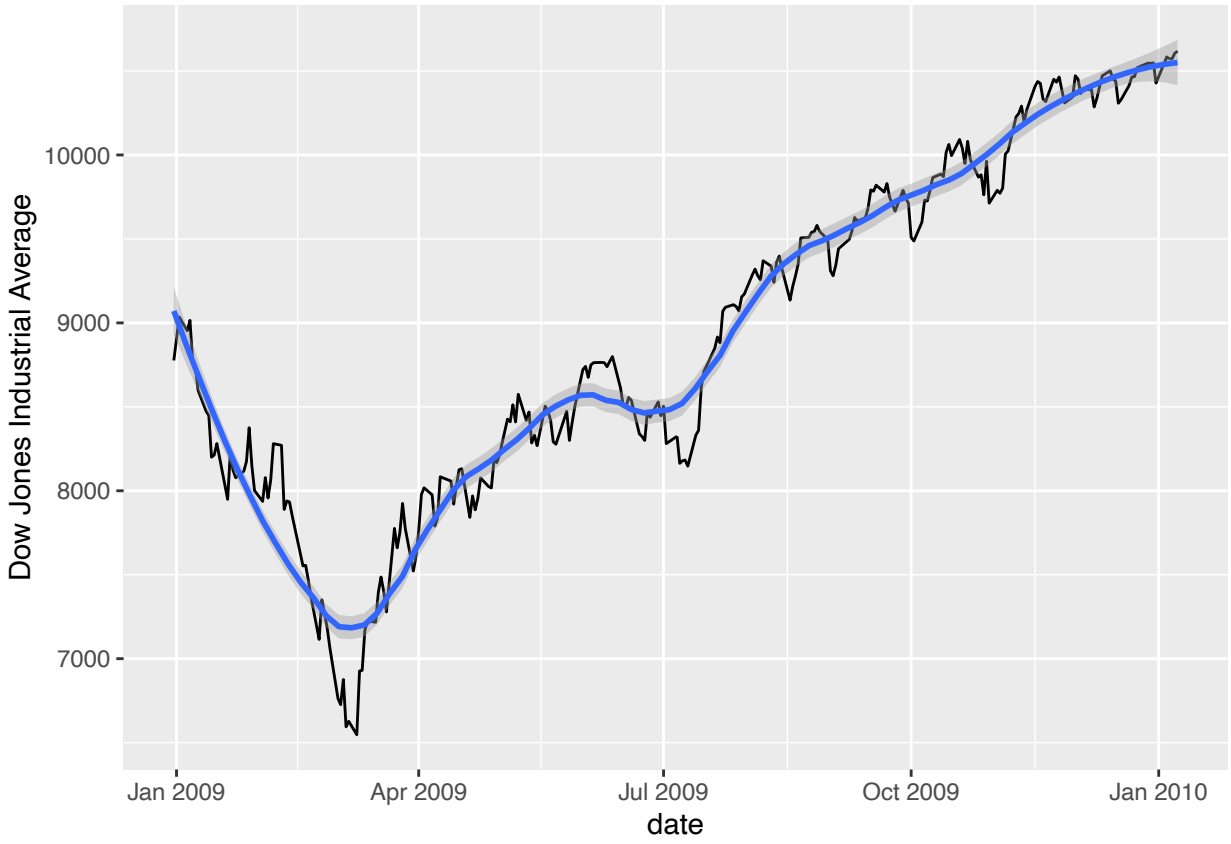
Plot 2 *Hints:*

- Wiggliness for `loess` is 0.3

```
ggplot(
  data = djia_date_range,
  mapping = aes(x = date, y = close)
) +
  geom_line() + geom_smooth(span = 0.3) +
  labs(
```

```
    y = "Dow Jones Industrial Average"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

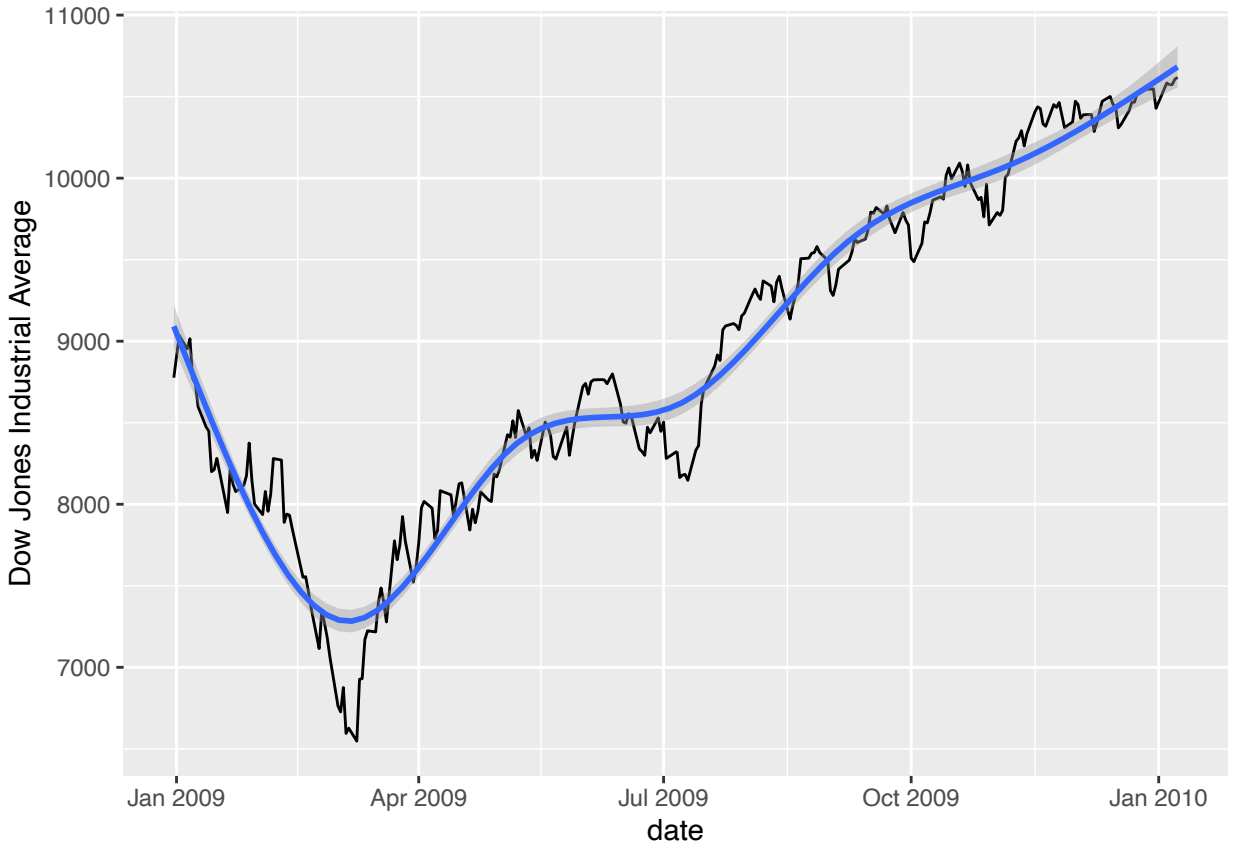


Plot 3 *Hints:*

- `y ~ ns(x, 6)` will need `splines` package (`"lm"` will work)

```
library(splines)

ggplot(
  data = djia_date_range,
  mapping = aes(x = date, y = close)
) +
  geom_line() + geom_smooth(method = "lm", formula = "y ~ ns(x, 6)") +
  labs(
    y = "Dow Jones Industrial Average"
  )
```



Exercise 3

Using `cdc` dataset, recreate the following graphics as precisely as possible.

```
# Read in the cdc dataset
cdc <- read_delim(file = "data/cdc.txt", delim = "|") %>%
  mutate(genhlth = factor(genhlth,
    levels = c("excellent", "very good", "good", "fair", "poor")
  ))
```

Plot 1 Construct this plot in two ways. Once using `cdc` and once using the `genhlth_count`.

```
genhlth_count <- cdc %>%
  count(genhlth)
```

##1. using cdc :

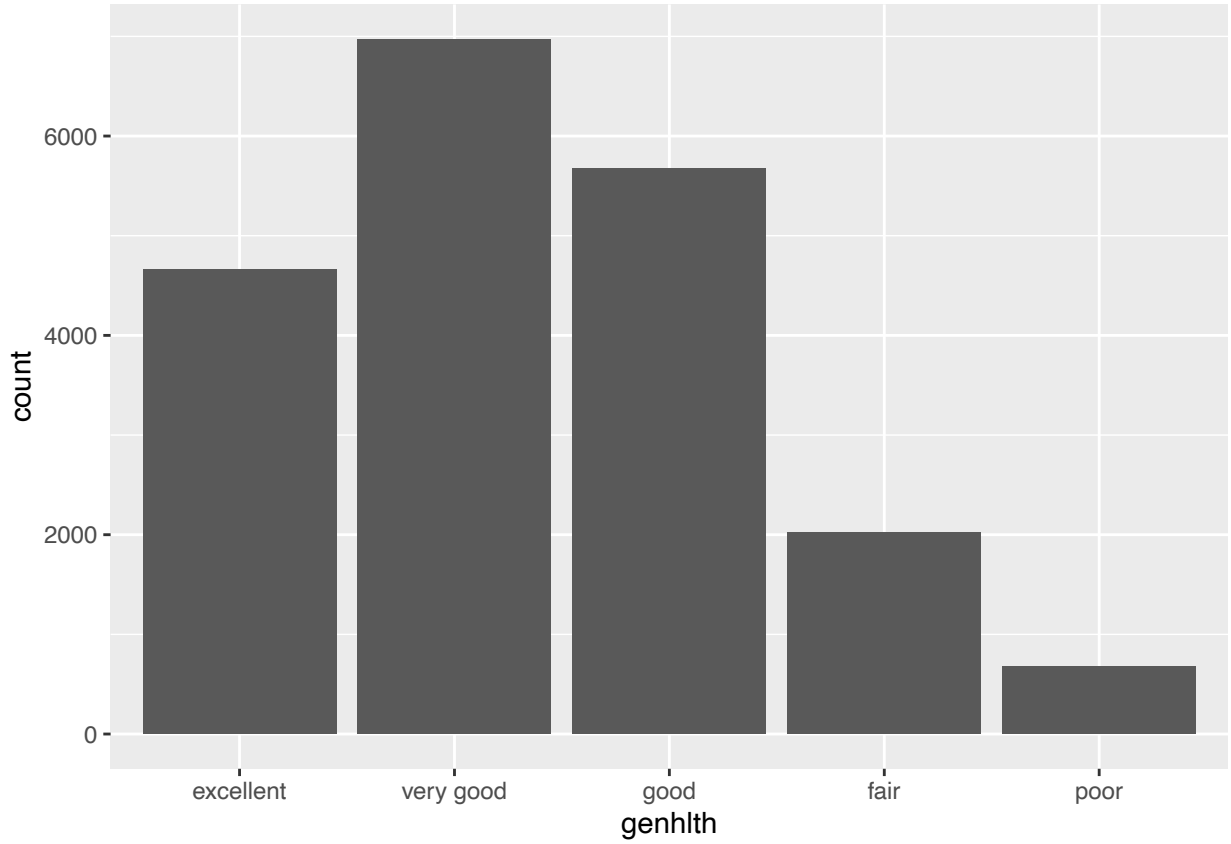
```
#make bar chart of genhlth vs count using cdc
cdc <- read_delim(file = "data/cdc.txt", delim = "|") %>%
  mutate(genhlth = factor(genhlth,
    levels = c("excellent", "very good", "good", "fair", "poor")
  ))
```

```
## Rows: 20000 Columns: 9
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "|"
## chr (2): genhlth, gender
## dbl (7): exerany, hlthplan, smoke100, height, weight, wtdesire, age
```

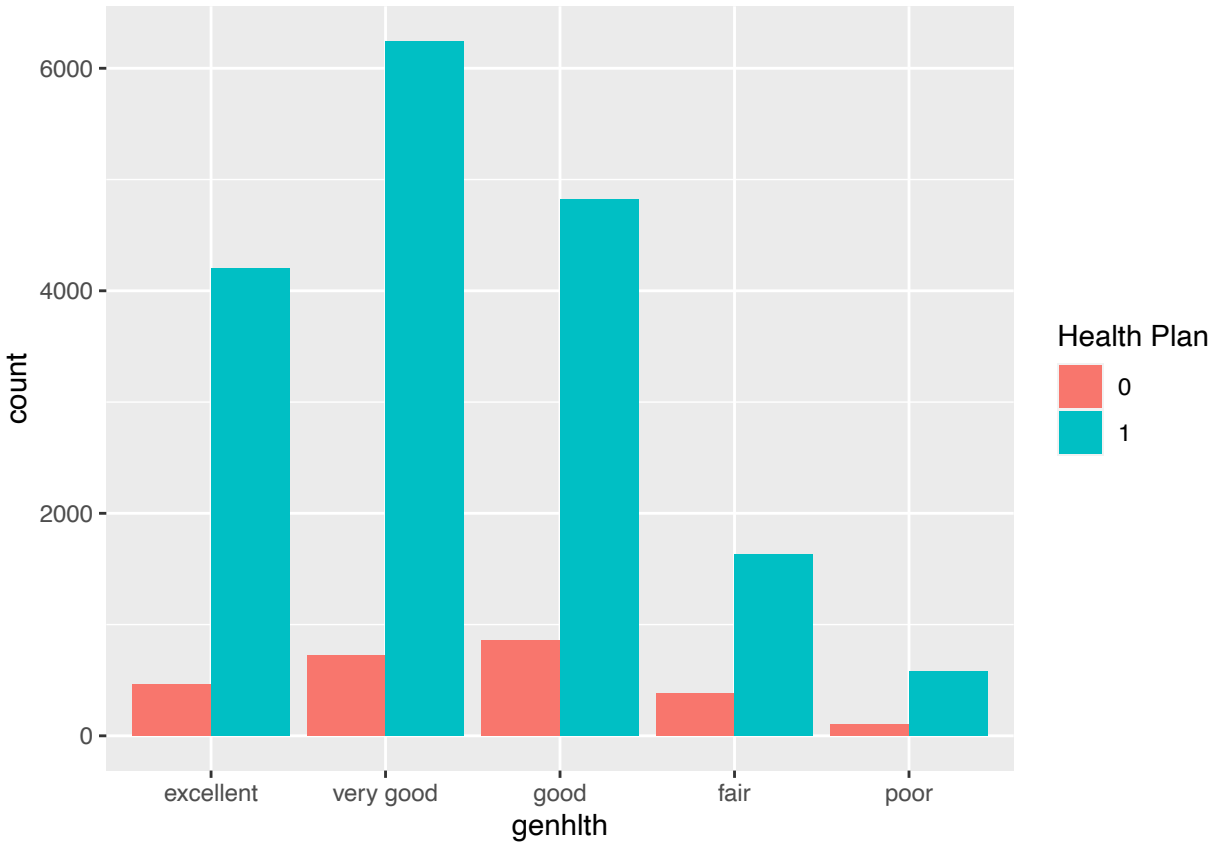
```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(data = cdc, aes(genhlth)) + geom_bar()
```



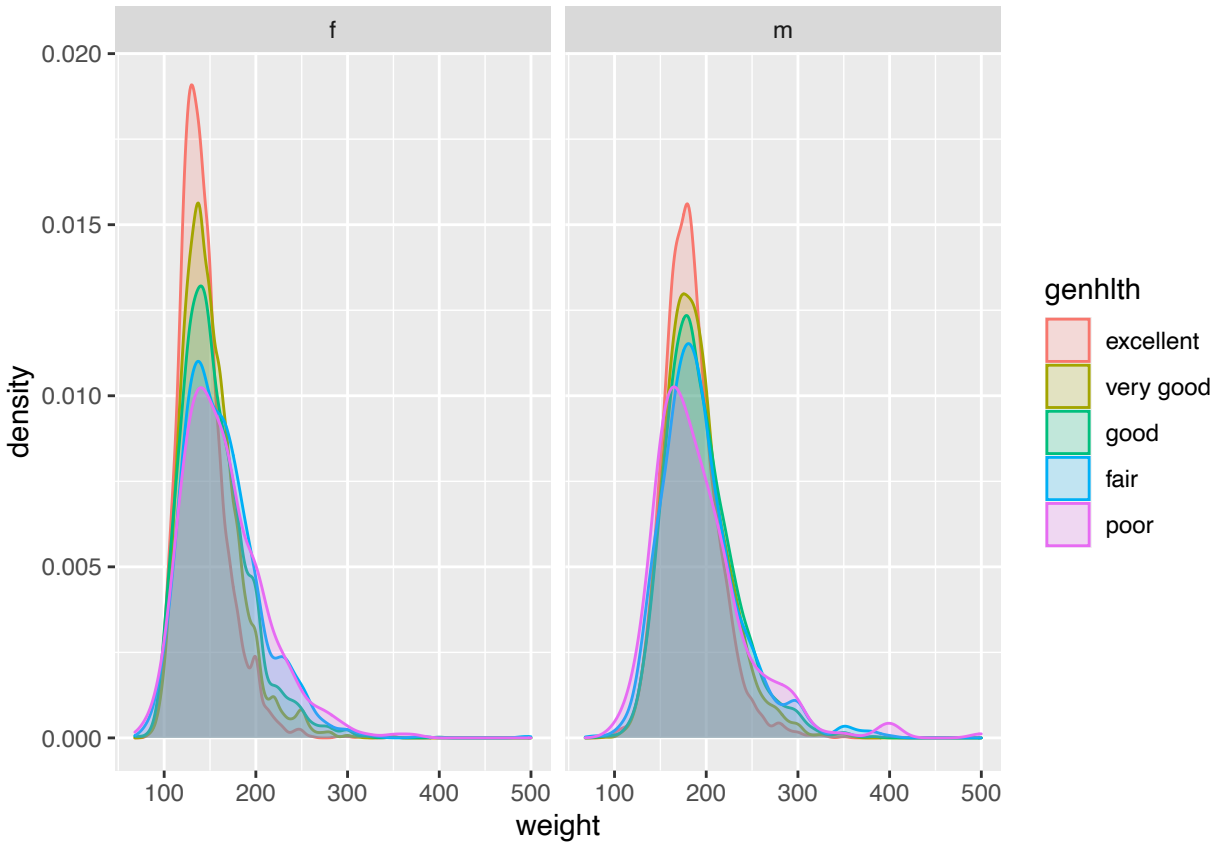

```
#make ggplot (bar chart) of genhlth vs count using cdc

ggplot(cdc, aes(genhlth, fill=as.factor(hlthplan))) +
  geom_bar(position="dodge") +
  labs(fill="Health Plan")
```



Plot 2

```
#plot density vs weight
ggplot(data = cdc, aes(weight, fill = genhlth, color = genhlth )) +
  geom_density(alpha = 0.2) +
  # facetwrap by sex
  facet_wrap(~gender)
```



Plot 3

Hints:

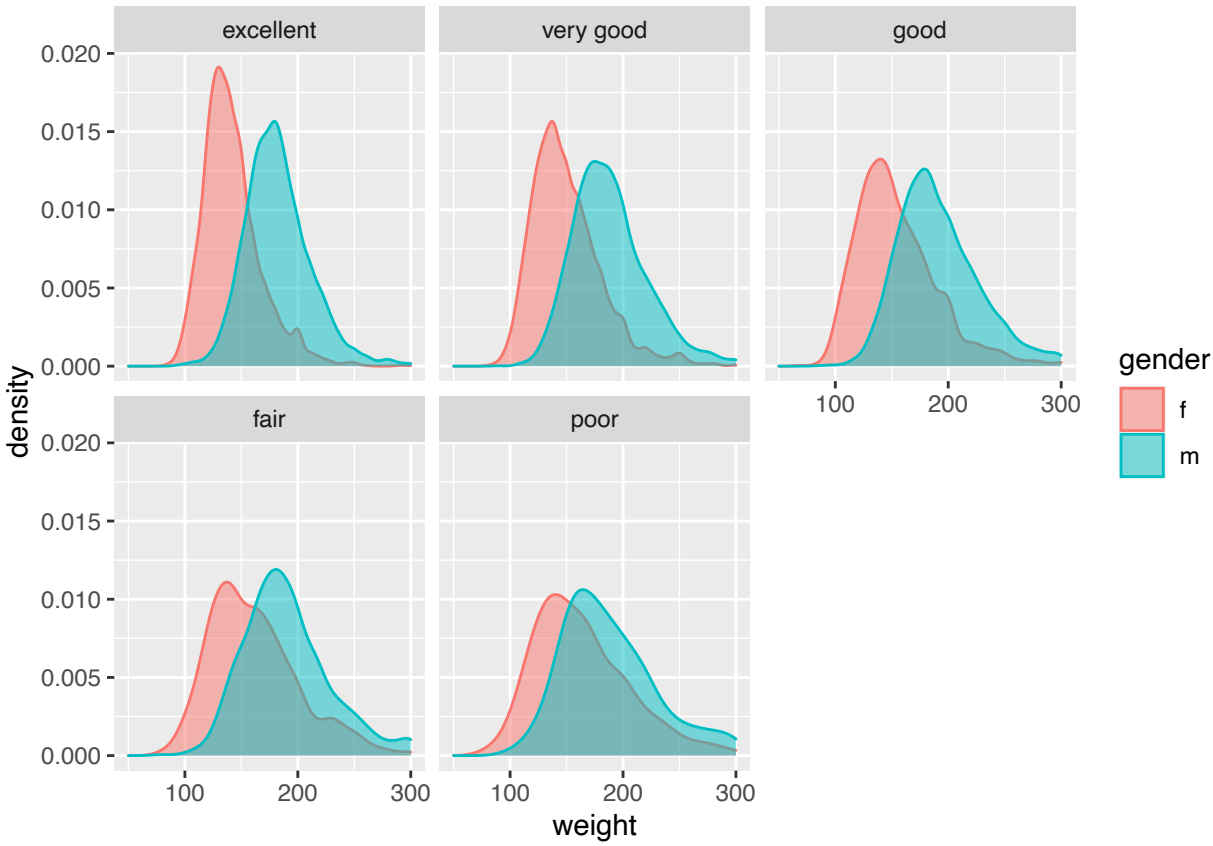
- Transparency is 0.2

Plot 4 *Hints:*

- Transparency is 0.5
- Horizontal axis should have lower limit of 50 and upper limit of 300

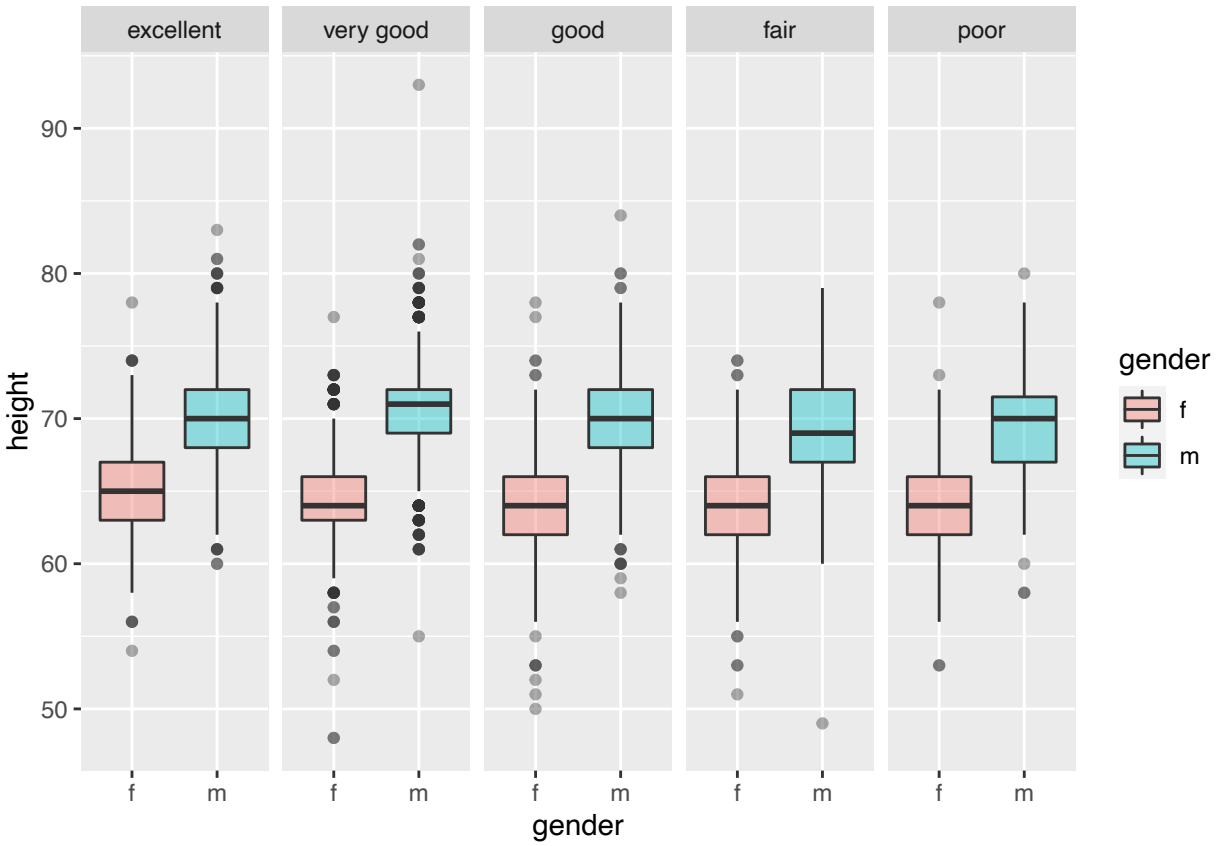
```
ggplot(data = cdc, aes(weight, fill = gender, color = gender)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~genhlth, scales = "fixed") +
  xlim(50, 300)
```

```
## Warning: Removed 103 rows containing non-finite values (stat_density).
```



Plot 5 *Hints:*

```
ggplot(data = cdc, aes(x = gender, y = height, fill = gender)) + geom_boxplot(alpha = 0.4) + facet_wrap
```

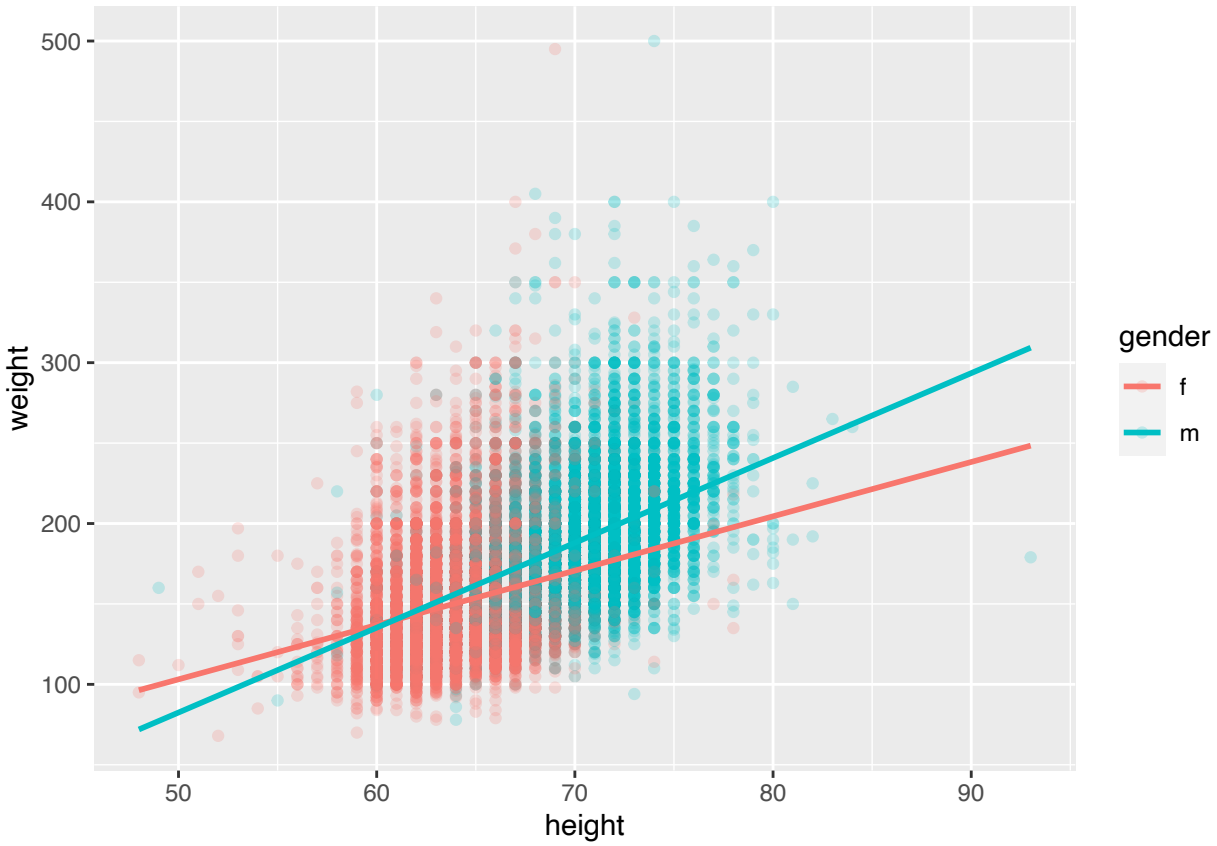


- Transparency is 0.4

Plot 6 *Hints:*

```
ggplot(data = cdc, aes(x = height, y = weight, color = gender)) +
  geom_point(alpha = 0.2) +
  geom_smooth(formula = y~x, se = FALSE, fullrange = TRUE)
```

```
## `geom_smooth()` using method = 'gam'
```



- Transparency is 0.2