Exercise: **ggplot Graphics**

Day 3, Part A

```
> library(ggplot2)
> library(RColorBrewer)
```

1. Load GBD2015 global deaths data ("data/gbd2015_global_deaths.csv"). These are estimates of the total number of deaths globally over time, including uncertainty intervals, from GBD. Discuss at your table:

   - How many rows and columns are there?
   - What are the classes (variable types) of each column?
   - What is the range of values for the numeric columns?
   - What are the possible values for the factor columns?
   - What does a single row represent?

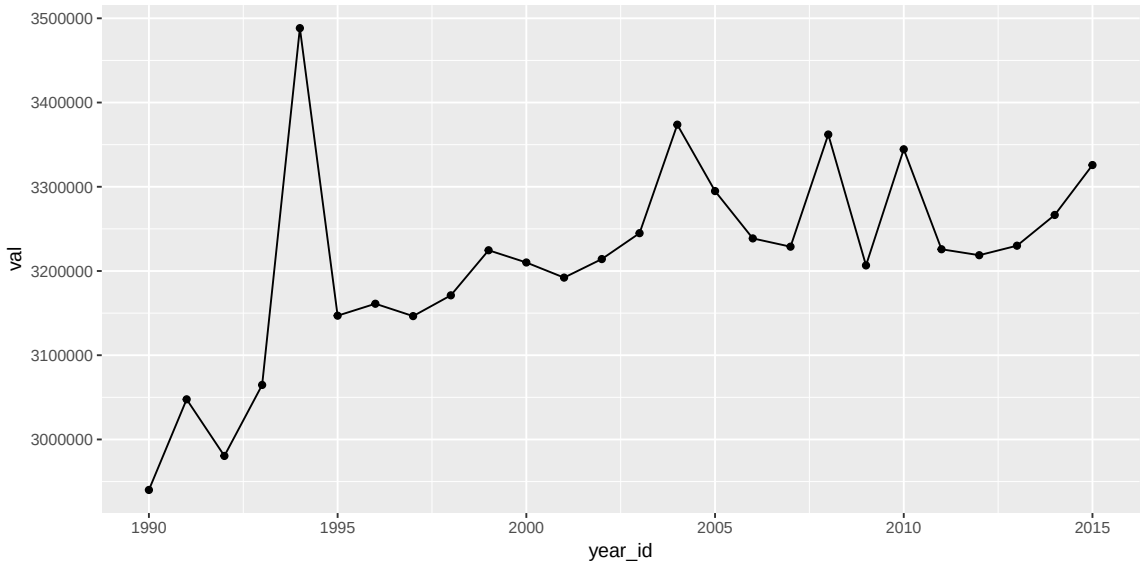
```
> main_dir <- "C:/Users/ngraetz/Documents/repos/r_training_penn/"  # CHANGE TO YOUR LOCAL COPY O
> data <- read.csv(paste0(main_dir, "data/gbd2015_global_deaths.csv"))
```

2. Subset the data to just rows that refer to injuries among males and store this as a new `data.frame`.

```
> subset_data <- data[data$sex == "Male" & data$cause_name == "Injuries", ]
```

   a. Use this subset of the data to make a graph of the number of deaths (the `val` column) over time like the one below:

```
> gg <- ggplot(subset_data,
+              aes(x = year_id,
+                  y = val)) +
+       geom_point() +
+       geom_line()
> gg
```



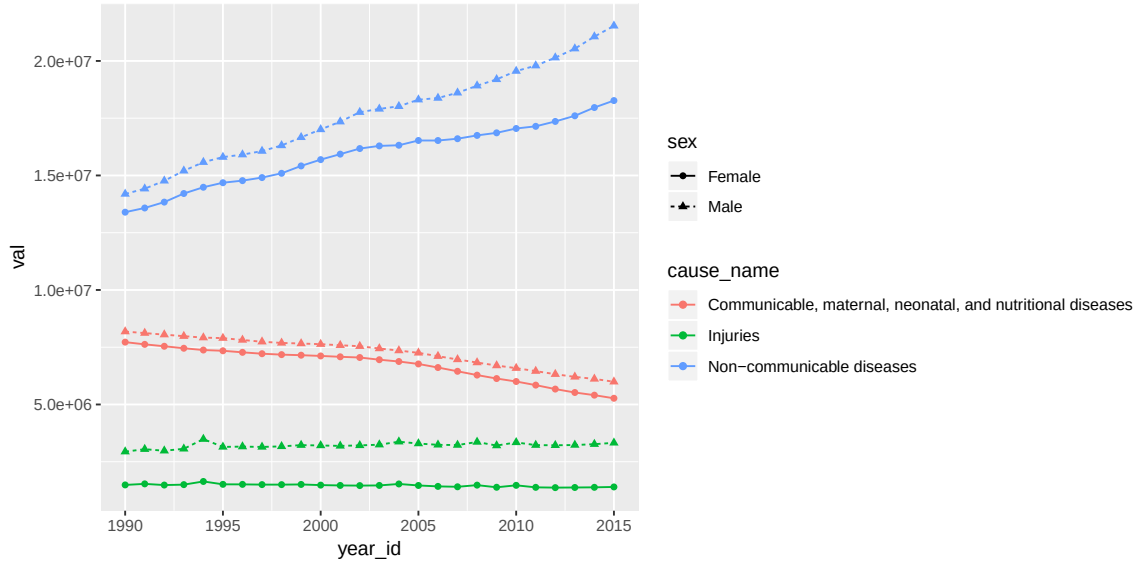
   b. Discuss at your table:

   - What are the `aesthetics` that need to be mapped to create this graph?
   - What are the `geoms` are involved in this graph?
   - What is the interpretation of this figure? Are there any notable years? What do you think might explain them?

3. Now using the full data,

   a. Recreate the graph below:

```
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val,
+                  color = cause_name,
+                  linetype = sex,
+                  shape = sex)) +
+       geom_point() +
+       geom_line()
> gg
```



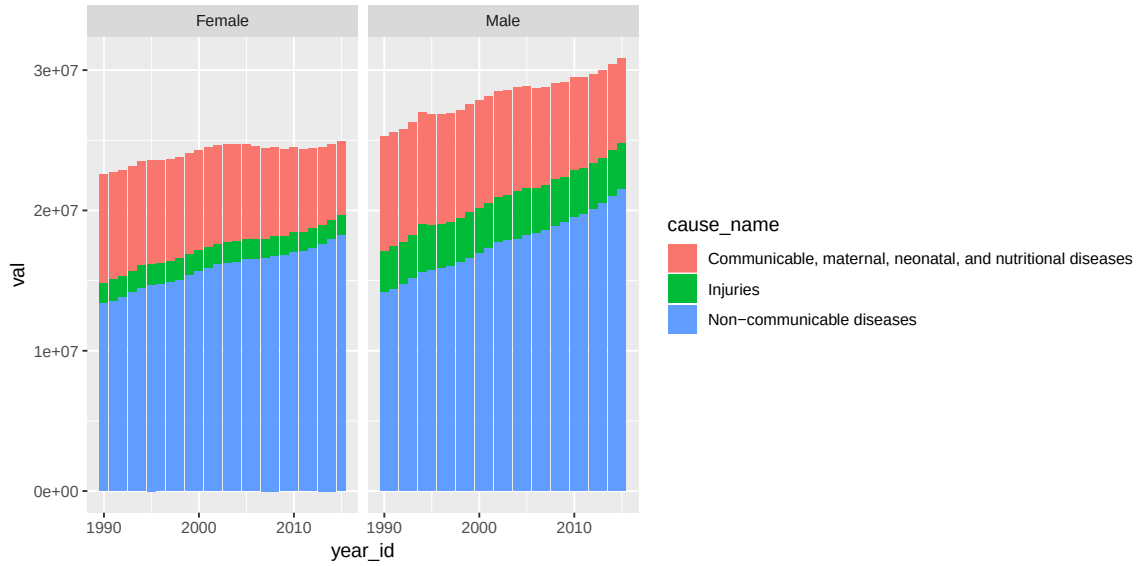
   b. Discuss at your table:

   - What are the `aesthetics` that need to be mapped to create this graph? (hint: there are three besides `x` and `y`, even though there are only two legends)
   - What are the `geoms` are involved in this graph?
   - What are the various ways you can interpret of this figure? What are the notable trends?

4. Using the same data,

   a. Recreate the bar graph below (hint: use `stat = "identity"` in the `geom`):

```
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val,
+                  fill = cause_name)) +
+       geom_bar(stat="identity") +
+       facet_wrap(~ sex)
> gg
```



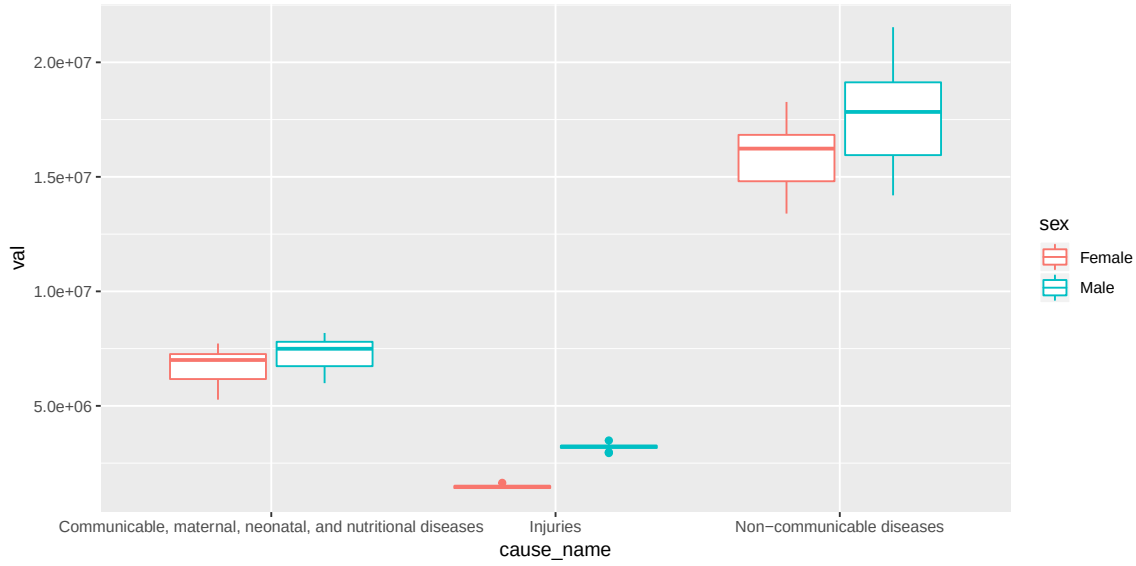
b. Discuss at your table:

- What are the `aesthetics` and `geoms` in this graph?
- What other variables appear in this graph? How?
- What would happen if you added `position='dodge'` to the `geom`? `position='fill'`?
- What are the various ways you can interpret of this figure? What are the notable trends?

5. Using the same data,

a. Recreate the box plot below:

```
> gg <- ggplot(data,
+              aes(x = cause_name,
+                  y = val,
+                  color = sex)) +
+       geom_boxplot()
> gg
```



b. Discuss at your table:

- What are the `aesthetics` and `geoms` in this graph?
- What are the various ways you can interpret of this figure? What is the range representing in each

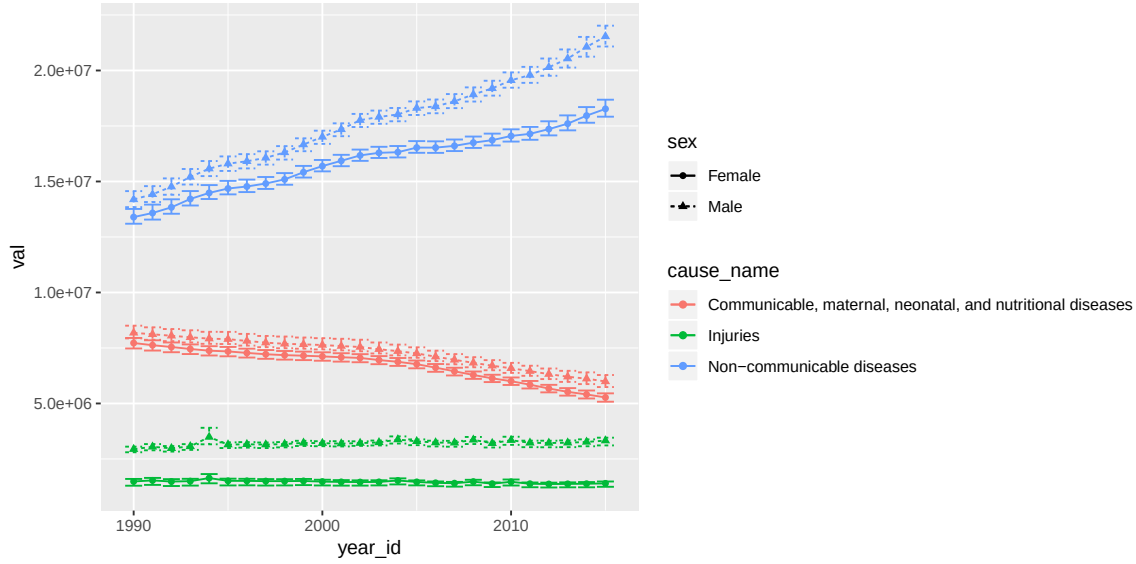
boxplot?

- If you wanted to look at the range across diseases by year (rather than the other way around), what `aesthetics` would you change? (hint: boxplots work best with factor variables on the x-axis)

6. Using the same data,

   a. Add error bars to the plot in question 3.

```
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val,
+                  ymin = lower,
+                  ymax = upper,
+                  color = cause_name,
+                  linetype = sex,
+                  shape = sex)) +
+       geom_point() +
+       geom_line() +
+       geom_errorbar()
> gg
```



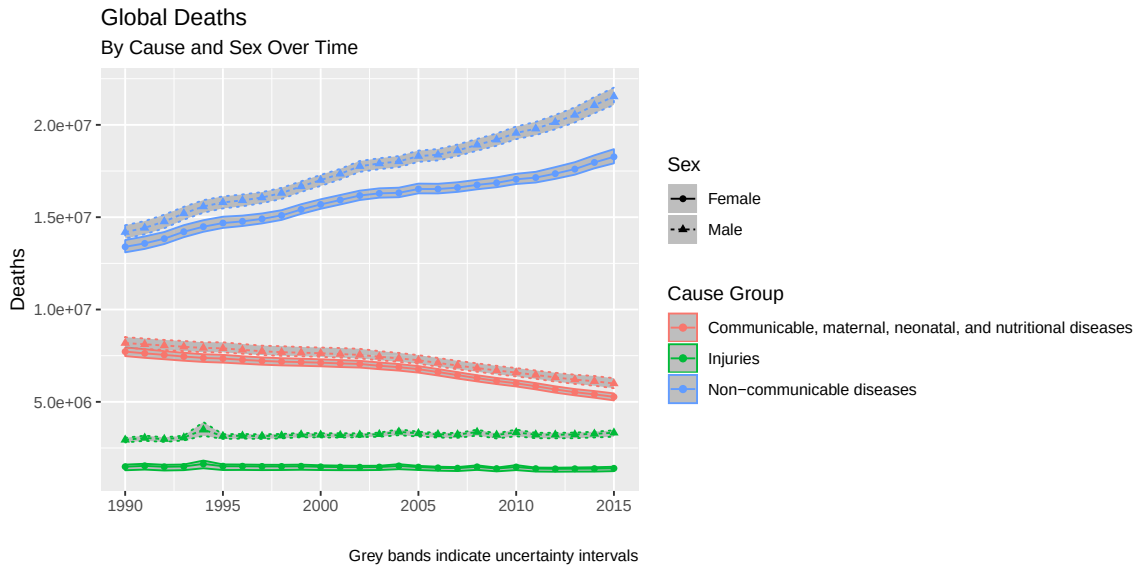
   b. Discuss at your table:

   - What are the `aesthetics` and `geoms` need to be added to this graph? (hint: look at the help file ?geom_errorbar to identify **required** aesthetics)
   - What are the trends in uncertainty? Are there particular cause groups, sexes or years with more or less uncertainty?

   c. Now try displaying uncertainty as a "ribbon", rather than bar:

```
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val,
+                  ymin = lower,
+                  ymax = upper,
+                  color = cause_name,
+                  linetype = sex,
+                  shape = sex)) +
+       geom_ribbon(fill='grey55', alpha=.5) +
+       geom_point() +
```

```
+       geom_line() +
+       labs(title='Global Deaths', subtitle='By Cause and Sex Over Time',
+         y='Deaths', x='', caption='Grey bands indicate uncertainty intervals',
+         shape='Sex', linetype='Sex', color='Cause Group')
> gg
```



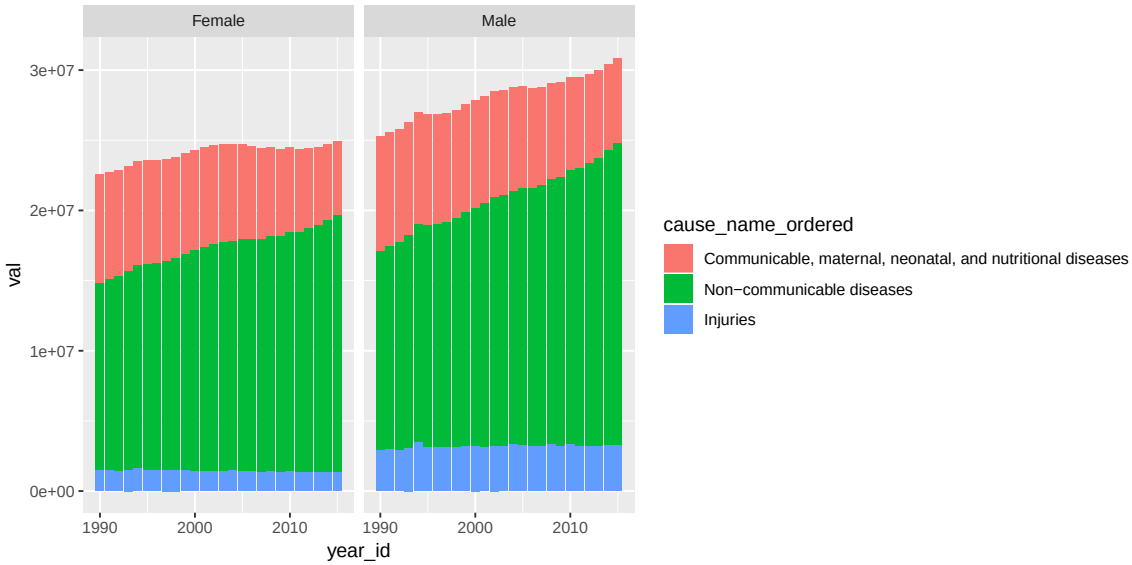
d. Discuss at your table:

- What order should the `geoms` appear to make sure everything is visible?
- How could you change the default color and transparency of the ribbon to make everything more visible?
- Does this alter any interpretations of uncertainty?
- Now that there are many things going on, I found it useful to label the graph better. What do you need to do to include all the labels I included? (notice that I also changed the legend titles)
- What happens if you give the `shape` aesthetic a different title than the `linetype` aesthetic?

Bonus Questions:

7. Change the order the causes are stacked in the plot in question 4 (communicable on the top, non-communicable in the middle, and injuries on the bottom).

```
> data$cause_name_ordered <- factor(data$cause_name,
+                                  levels = c("Communicable, maternal, neonatal, and nutritiona
+                                             "Non-communicable diseases",
+                                             "Injuries"))
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val,
+                  fill = cause_name_ordered)) +
+       geom_bar(stat="identity") +
+       facet_wrap(~ sex)
> gg
```

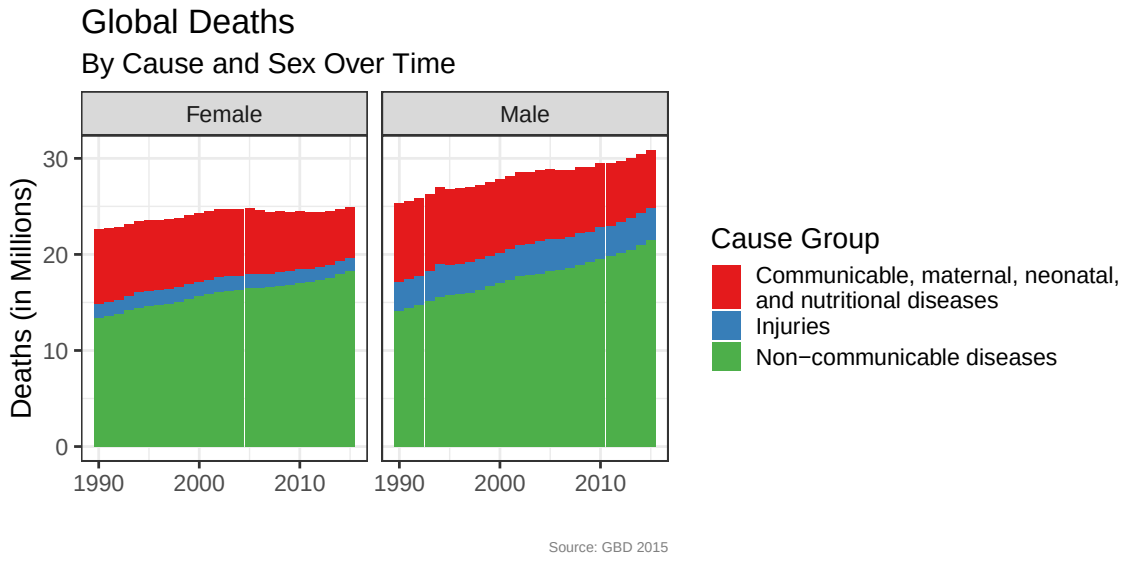


Hint: this is a general R question, not a ggplot2 question

8. Make a "production ready" version of the stacked bar graph above (one that would be clean enough to publish)

   - Use `brewer.pal` (from the `RColorBrewer` package), or choose your own colors to assign nicer colors
   - Use `themes` to format the background and enlarge the text
   - Use `labs` to provide better titles
   - Use "on-the-fly" data manipulation to avoid scientific notation on the y-axis
   - Use `str_wrap()` from the `stringr` package to wrap the text in the legend
   - Use `theme(plot.caption=element_text(size=8, color='grey50'))` to modify the caption text separately

```
> library(stringr)
> gg <- ggplot(data, aes(x = year_id, y = val/1e+06, fill = str_wrap(cause_name, 35))) +
+     geom_bar(stat = "identity") + facet_wrap(~sex) + scale_fill_manual(values = brewer.pal(3,
+     "Set1")) + labs(title = "Global Deaths", subtitle = "By Cause and Sex Over Time",
+     y = "Deaths (in Millions)", x = "", fill = "Cause Group", caption = "Source: GBD 2015") +
+     theme_bw(base_size = 16) + theme(plot.caption = element_text(size = 8, color = "grey50"))
> gg
```



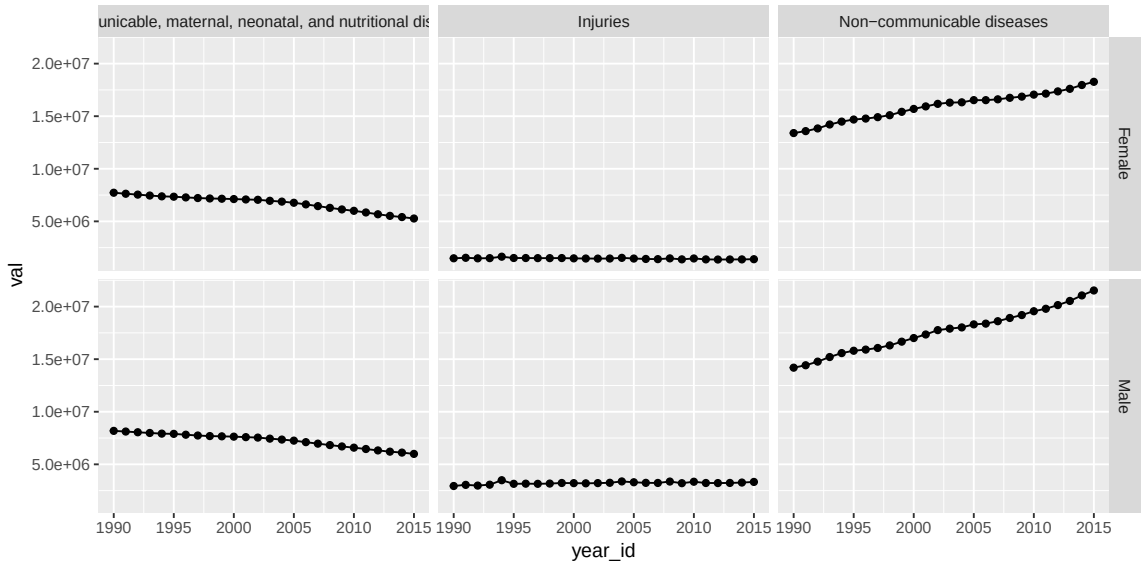
Hints:

- https://ggplot2.tidyverse.org/reference/scale_manual.html
- https://ggplot2.tidyverse.org/reference/theme.html
- http://colorbrewer2.org/

9. Explore more complex ways of creating `facets`.

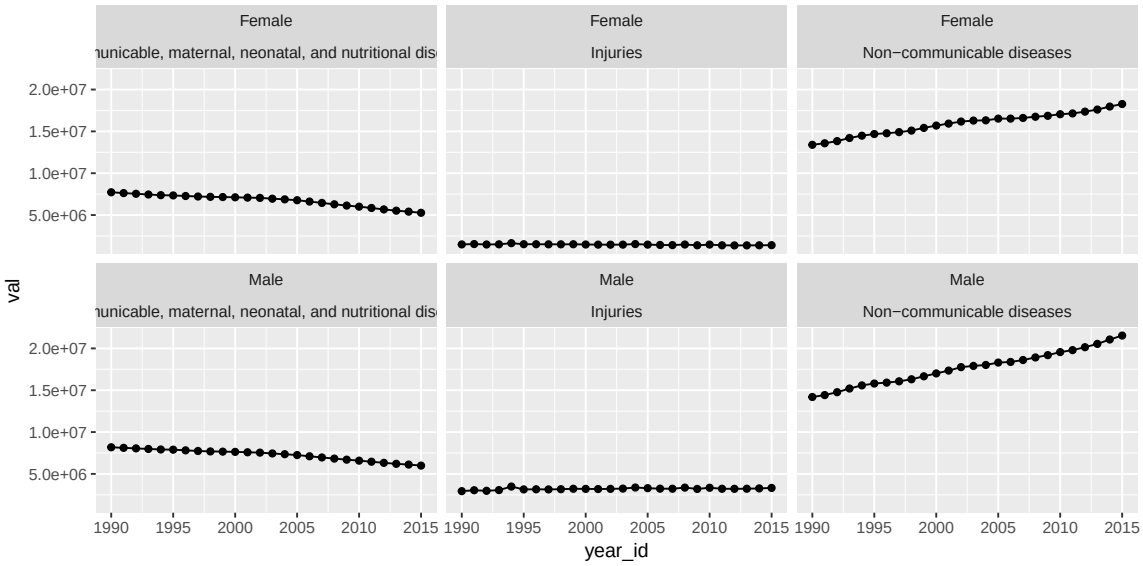
   a. Can you figure out how to make a graph like the below?

```
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val)) +
+       geom_point() +
+       geom_line() +
+       facet_grid(sex~cause_name)
> gg
```



   b. Or this one?

```
> gg <- ggplot(data,
+              aes(x = year_id,
+                  y = val)) +
+       geom_point() +
+       geom_line() +
+       facet_wrap(~sex+cause_name)
> gg
```


Female
unicable, maternal, neonatal, and nutritional dis
Female
Injuries
Female
Non−communicable diseases
Male
unicable, maternal, neonatal, and nutritional dis
Male
Injuries
Male
Non−communicable diseases
val
2.0e+07
1.5e+07
1.0e+07
5.0e+06
1990
1995
2000
2005
2010
2015
year_id