Assignment 4: Mind the Gap

Minh Tran

2021-05-27

Exercise 1

i. The variables that are categorical are 'country', 'continent', and 'region'.

ii. The variables that are continous are 'infant_mortality', 'life_expectancy', and 'fertility'.

iii. Each row represents the health and income outcomes of a country.

Exercise 2

i.

```
gapminder %>%
  group_by(continent) %>%
  summarize(
    mean = mean(fertility,na.rm = TRUE),
    median = median(fertility,na.rm = TRUE),
    std_dev = sd(fertility, na.rm = TRUE),
    iqr = IQR(fertility, na.rm = TRUE),
    min = min(fertility, na.rm = TRUE),
    max = max(fertility, na.rm = TRUE),
    numobs = n(),
    missobs = sum(is.na(fertility)),
    fractionmiss = missobs/numobs
  )
```

continent	mean	median	std_dev	iqr	min	max	numobs	missobs	fractionmiss
Africa	5.850994	6.160	1.4051802	1.7000	1.50	8.45	2907	51	0.0175439
Americas	3.652681	3.140	1.5918369	2.4500	1.45	7.56	2052	38	0.0185185
Asia	4.181983	3.855	1.9654074	3.5000	0.84	9.22	2679	47	0.0175439
Europe	1.967500	1.870	0.6124244	0.7225	1.13	6.19	2223	39	0.0175439
Oceania	4.354420	4.390	1.5907184	2.4800	1.73	7.65	684	12	0.0175439

ii.

```
gapminder %>%
  group_by(continent) %>%
  summarize(
    numobs = n(),
    missobs = sum(is.na(region)),
    fractionmiss = missobs/numobs
  )
```

continent	numobs	missobs	fractionmiss
Africa	2907	0	0
Americas	2052	0	0
Asia	2679	0	0
Europe	2223	0	0
Oceania	684	0	0

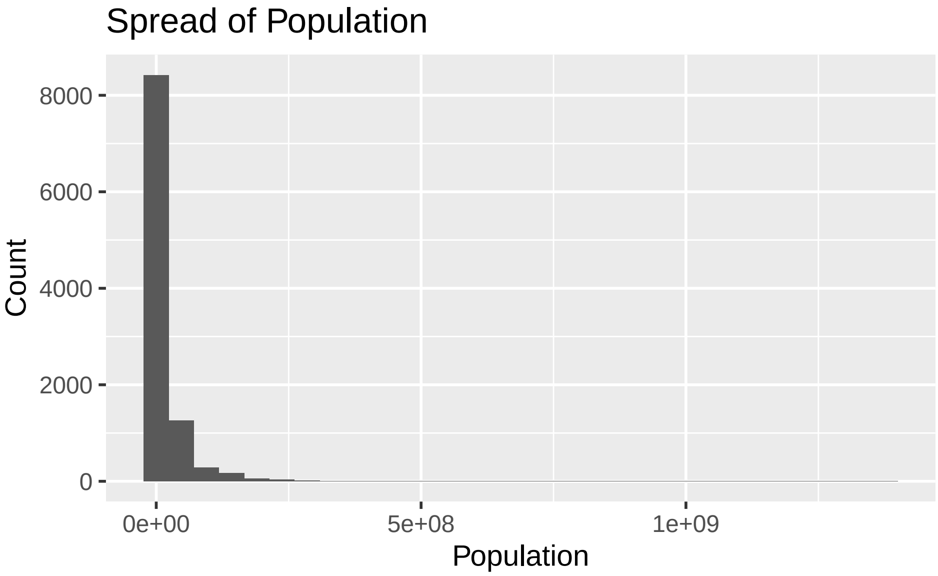
Exercise 3

i.

```
gapminder %>%
  ggplot() +
  geom_histogram(aes(x = population)) +
  labs(title = 'Spread of Population') +
  xlab('Population') +
  ylab('Count')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 185 rows containing non-finite values (stat_bin).
```

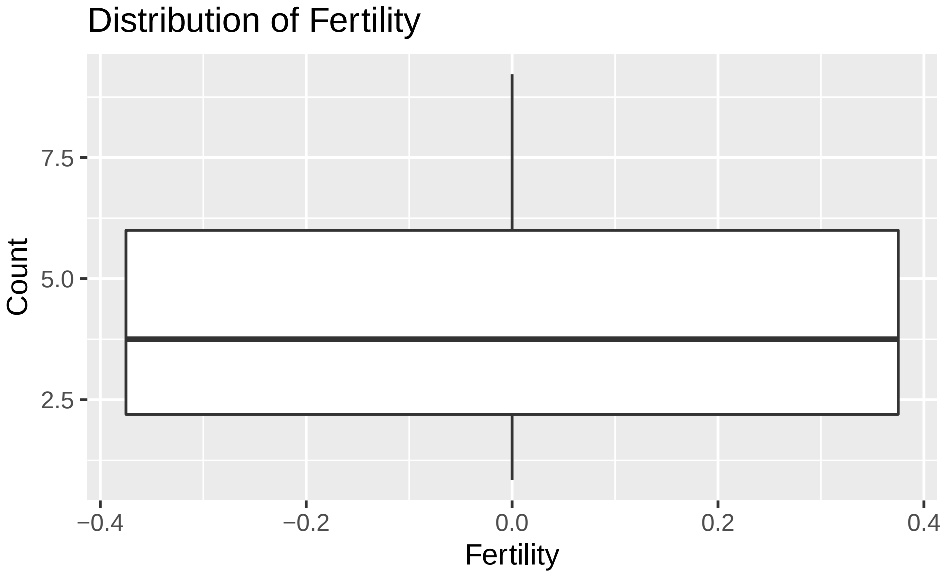


*** The shape of the histogram is unimodal(right skewed).Majority of the countries have a population in that interval. ***

ii.

```r
gapminder %>%
  ggplot() +
  geom_boxplot(aes(y = fertility)) +
  labs(title = 'Distribution of Fertility') +
  xlab('Fertility') +
  ylab('Count')
```

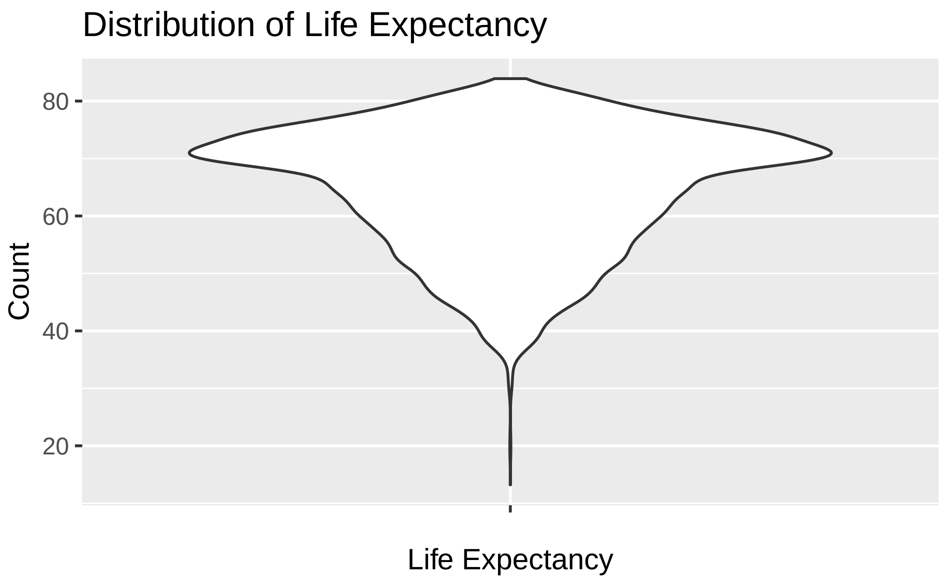
```
## Warning: Removed 187 rows containing non-finite values (stat_boxplot).
```



Center of distribution is between 2.5 and 5.0 (around 3.75).

iii.

```r
gapminder %>%
  ggplot() +
  geom_violin(aes(x = "", y = life_expectancy)) +
  labs(title = 'Distribution of Life Expectancy') +
  xlab('Life Expectancy') +
  ylab('Count')
```

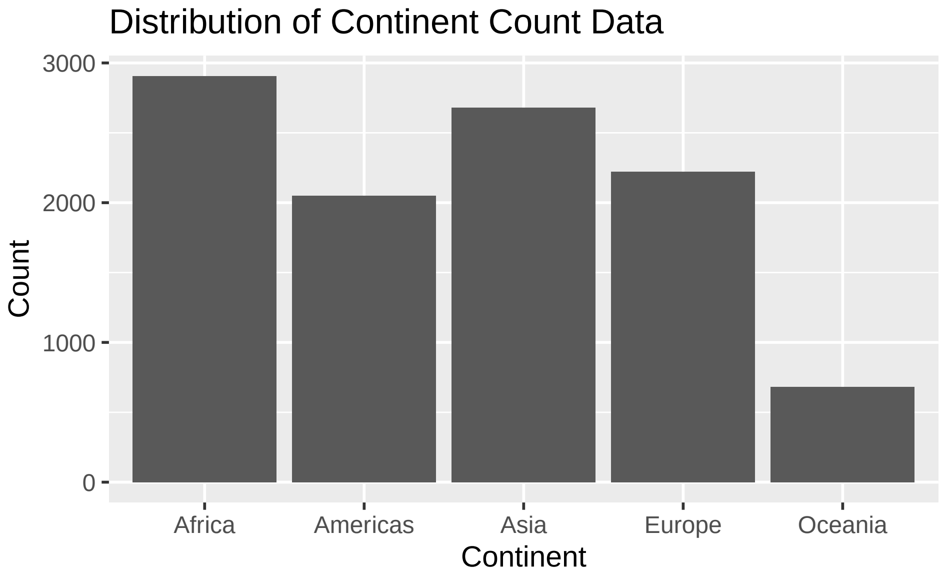


Looks unimodal with a center of distribution between 60 and 80 (around 70).

Exercise 4

i.

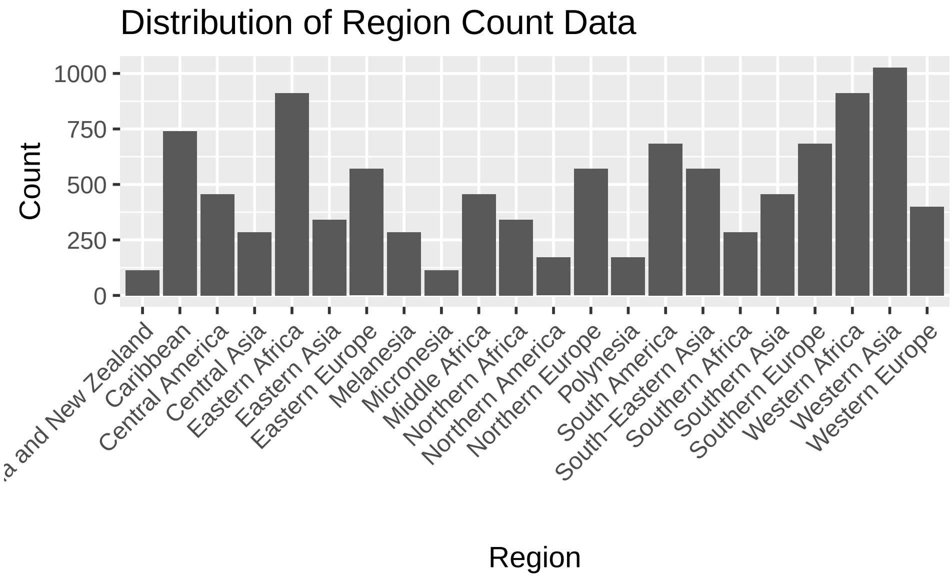
```r
gapminder %>%
  ggplot() +
  geom_bar(aes(x = continent)) +
  labs(title = 'Distribution of Continent Count Data') +
  xlab('Continent') +
  ylab('Count')
```



Since each row is representing a country, more than one country are going to belong to a continent, so that is why there are many counts of the continents.

ii.

```
gapminder %>%
  ggplot() +
  geom_bar(aes(x = region)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = 'Distribution of Region Count Data') +
  xlab('Region') +
  ylab('Count')
```



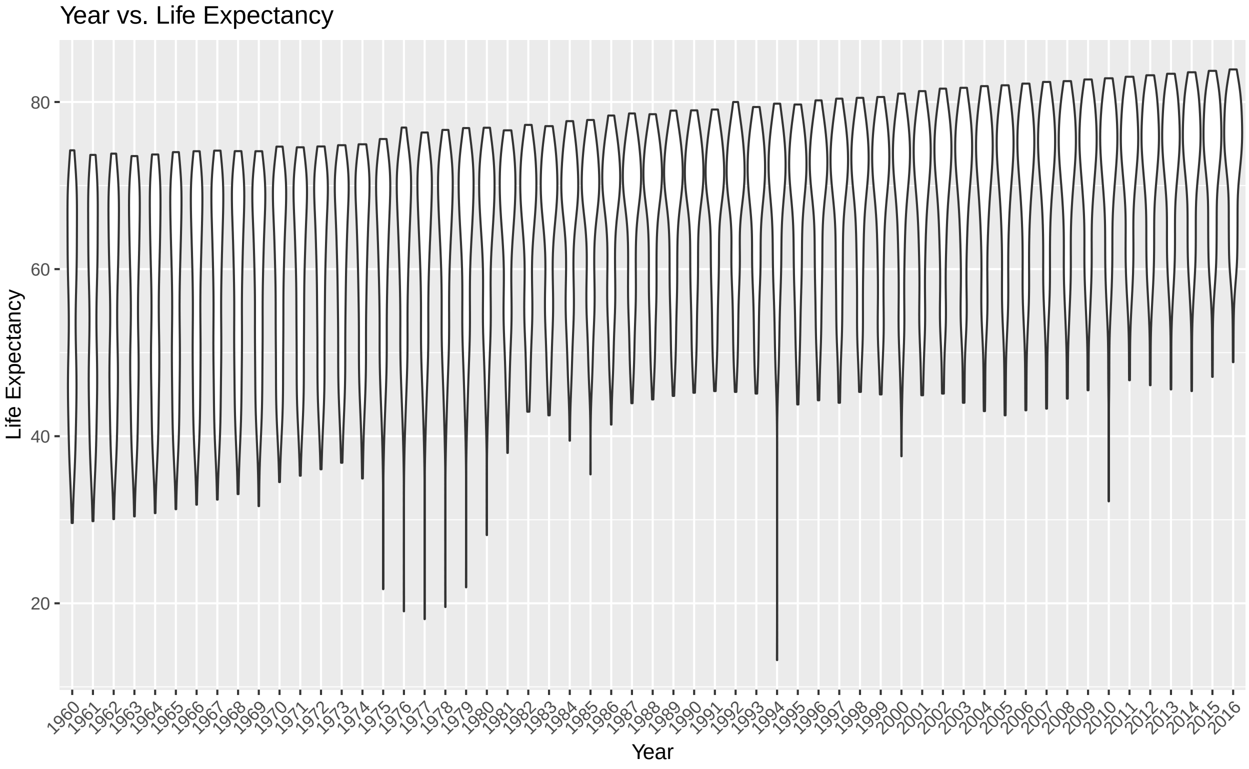
Exercise 5

i.

```
gapminder_cat <- gapminder %>%
  mutate(year_cat = as.factor(year))
```

ii.

```
gapminder_cat %>%
  ggplot() +
  geom_violin(aes(x = year_cat, y = life_expectancy)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(title = 'Year vs. Life Expectancy') +
  xlab('Year') +
  ylab('Life Expectancy')
```



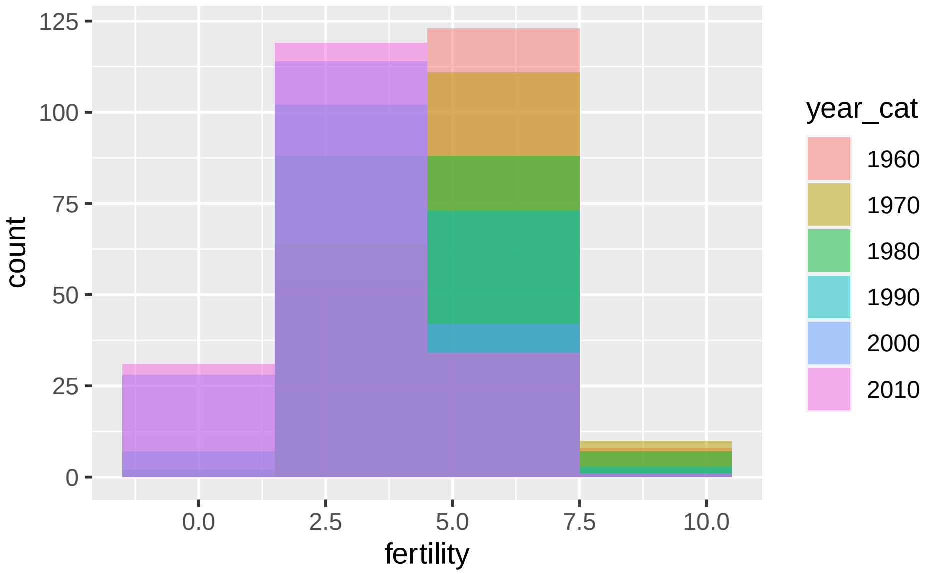
iii. ***It is showing that the global trend of life expectancy over time is increasing.***

iv. ***The center of distribution of life expectancy has increased over time (unimodal). People are living longer lives over time.***

Exercise 6

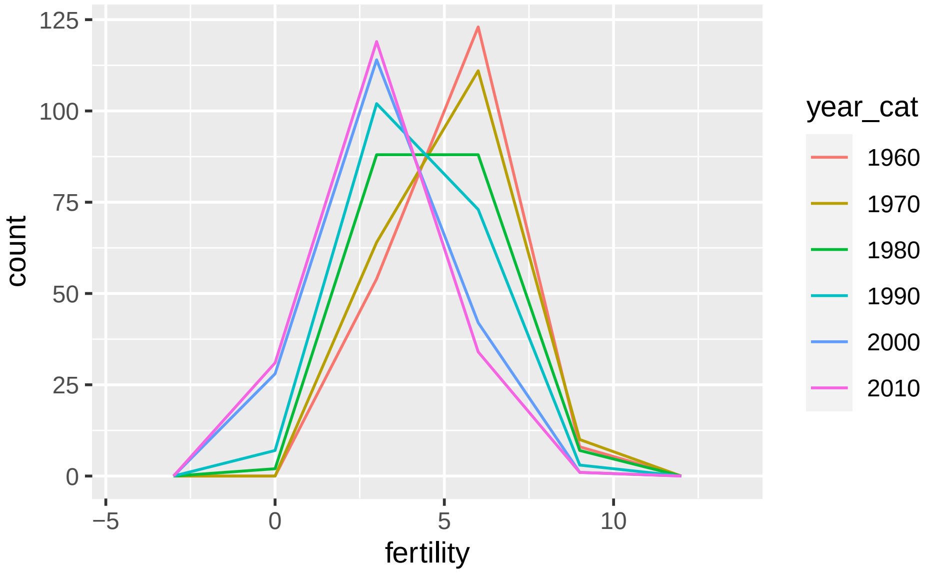
i.

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
  geom_histogram(aes(x = fertility, fill = year_cat), binwidth = 3,
  position = 'identity', alpha = 0.5)
```



ii.

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
  geom_freqpoly(aes(x = fertility, color = year_cat), binwidth = 3)
```

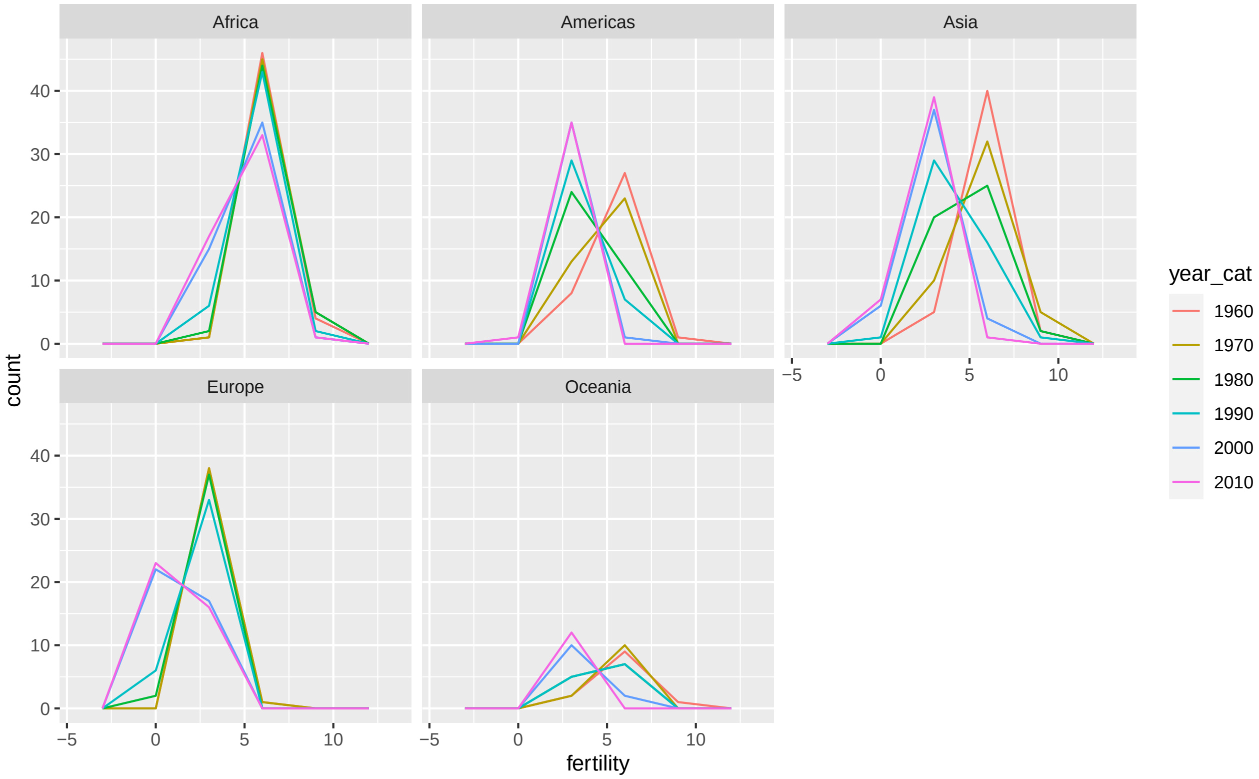


iii. ***The center of distribution for fertility has decreased over time. Over the years, the average number of children per woman decreased.***

Exercise 7

i.

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
  geom_freqpoly(aes(x = fertility, color = year_cat), binwidth = 3) +
  facet_wrap(~ continent)
```



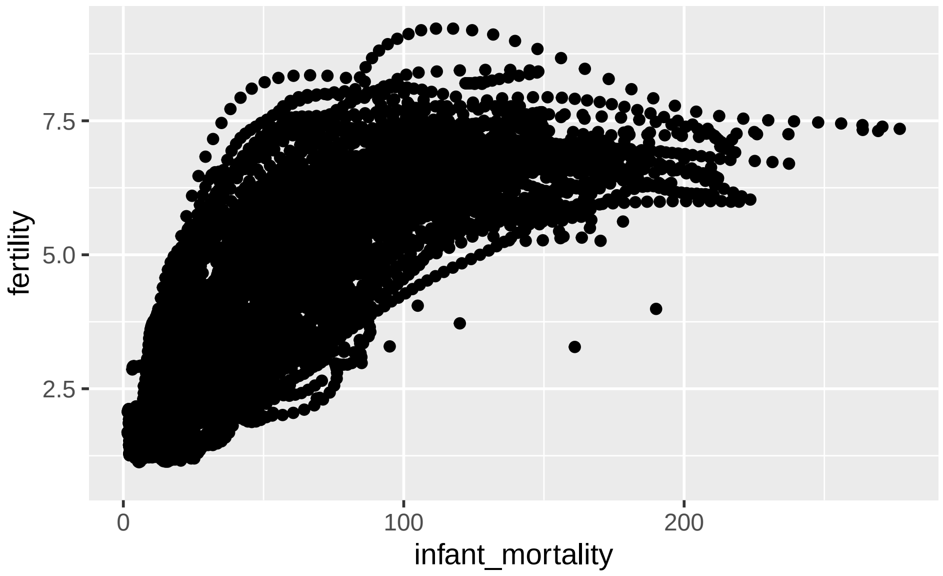
ii. *There are definitely differences in how fertility has changed over time in different continents. For instance, the fertility rate for Africa and Europe stayed consistent for many years compared to the other continents. Contienents like Africa seem to not have a huge decline in fertility rates compared to other continents. Overall, most continents are similar to each other by 2010 by which the fertility rates have dropped globally.*

Exercise 8

i.

```
gapminder %>%
  ggplot() +
  geom_point(aes(x = infant_mortality, y = fertility))
```

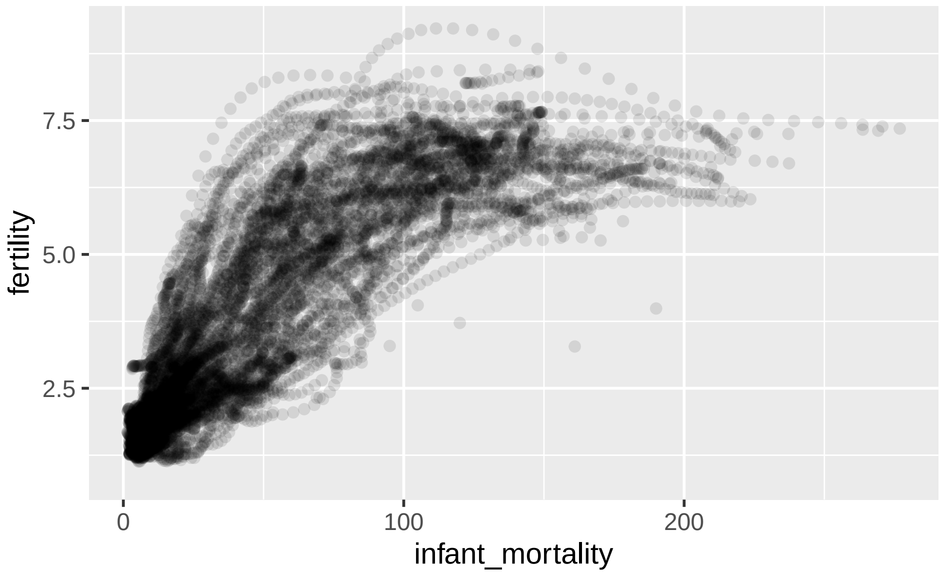
```
## Warning: Removed 1453 rows containing missing values (geom_point).
```



ii.

```
gapminder %>%
  ggplot() +
  geom_point(aes(x = infant_mortality, y = fertility), alpha = 0.1)
```

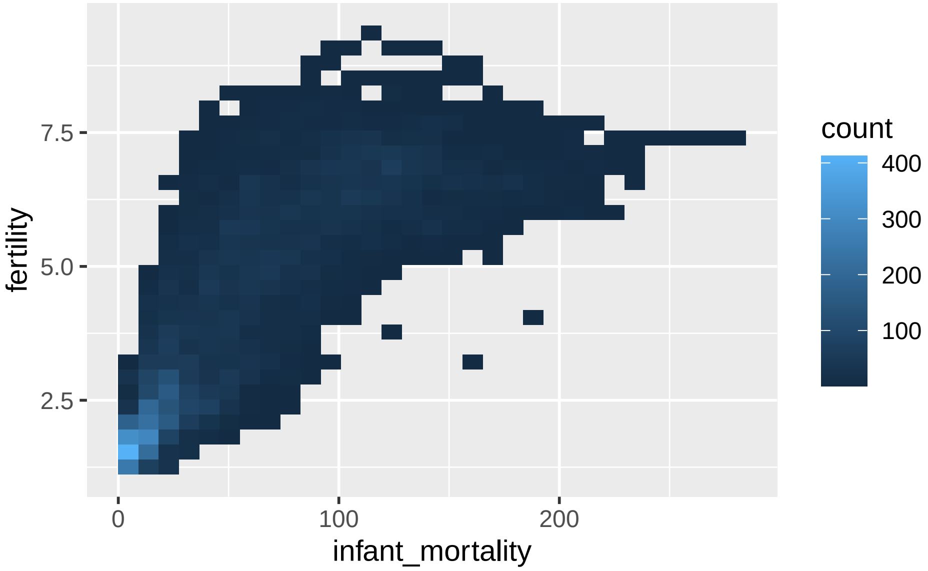
```
## Warning: Removed 1453 rows containing missing values (geom_point).
```



iii.

```
gapminder %>%
  ggplot() +
  geom_bin2d(aes(x = infant_mortality, y = fertility))
```

```
## Warning: Removed 1453 rows containing non-finite values (stat_bin2d).
```

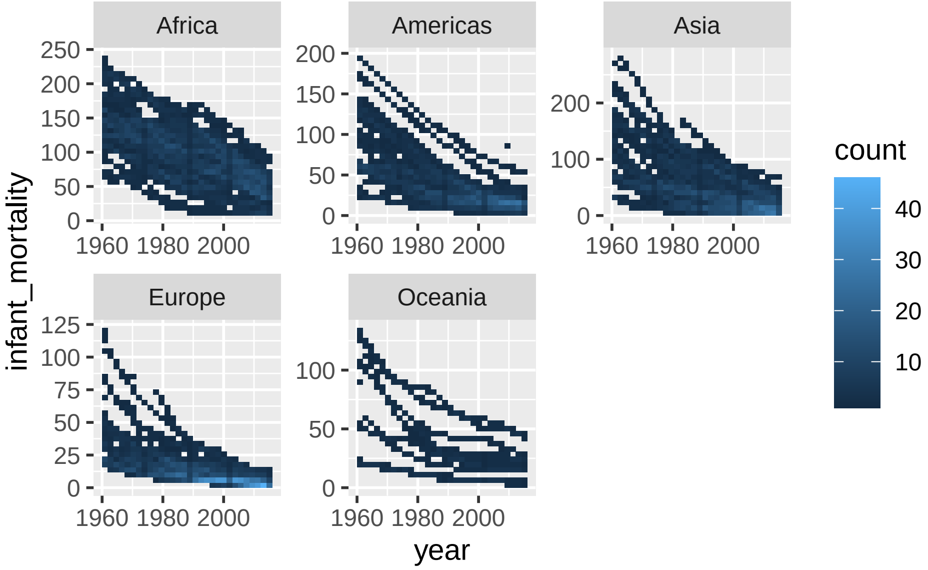


iv. *There are higher infant_morality rates as the fertility rates increases. I think both graphs are very similar and show the same information such as the relationship between infant morality and fertility as well as the where most of the data points are. However, the heatmap looks more visually appealing than the scatterplot.*

Exercise 9

```
gapminder %>%
  ggplot() +
  geom_bin2d(aes(x = year, y = infant_mortality)) +
  facet_wrap(~continent, scales = 'free')
```

```
## Warning: Removed 1453 rows containing non-finite values (stat_bin2d).
```



It looks like the infant mortality rate has decreased over the years for all continents. Less infant deaths can be linked to modern health and medicine which also contributed to longer lifespans. Most of the datapoints are around the 2000s. By the 2000s, the continents are similar to each other in terms of the infant mortality rate decreasing substantially.